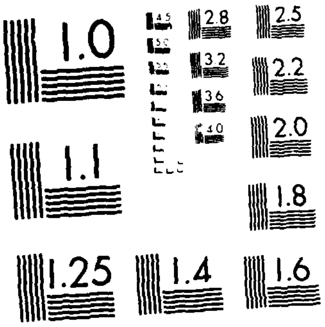
AD-A194 163 PERFORMANCE OF STOCHASTIC AND DECENTRALIZED SYSTEMS(U) 1/2
CALIFORNIA UNIV SANTA BARBARA DEPT OF ELECTRICAL AND
COMPUTER ENGINEERING I B RHODES 11 FEB 88
UNCLASSIFIED N00014-59-C-0459 F/G 25/5 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0

AD-A194 163

Unclassified

4

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
PERFORMANCE OF STOCHASTIC AND DECENTRALIZED SYSTEMS

5. TYPE OF REPORT & PERIOD COVERED
FINAL: 1 JUL 79 - 30 JUN 85

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Ian B. Rhodes

8. CONTRACT OR GRANT NUMBER(s)
N00014-59-C-0459

9. PERFORMING ORGANIZATION NAME AND ADDRESS
University of California, Santa Barbara
Santa Barbara, CA 93106

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 041-500 (410)

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research Code 432
800 N. Quincy Street
Arlington, VA 22217

12. REPORT DATE
11 FEB 88

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Office of Naval Research Detachment, Pasadena
1030 East Green Street
Pasadena, CA 91106

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stochastic systems, stochastic processes, point-processes, Poisson processes, optical communication, cutoff rate, shortest path problems, decentralized control, networks, decentralized estimation, decentralized detection, filtering, smoothing, likelihood ratio tests, estimators, observers, minimum-order observers, decentralized decision making

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Included in this report are results of investigations into the following topics: optimal signal design for coded direct-detection optical communication systems; informationally-decentralized network problems; decentralized estimation and control; nonlinear smoothing algorithms; minimum-order observers and state estimators; decentralized sequential detection and decision-making.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

PERFORMANCE OF STOCHASTIC AND DECENTRALIZED SYSTEMS

FINAL REPORT

CONTRACT N00014-59-C-0459

Ian B. Rhodes
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106

February 11, 1988

TABLE OF CONTENTS

Introduction 1

Accomplishments 2

Optimum Signal Design for Coded, Direct-Detection Optical Communication Systems. 2

Informationally-Decentralized Network Problems. 3

Decentralized Estimation and Control 4

Nonlinear Smoothing Algorithms. 5

Minimum-Order Observers and State Estimators. 6

Decentralized Sequential Detection and Decision Making. 9

Theses Supported 11

List of Publications 12

Appendix 14

FINAL REPORT

CONTRACT N00014-59-C-0459

I. INTRODUCTION

The research performed under this Contract has been concerned with a number of topics within the general area of decision, estimation and control for systems operating in an environment of uncertainty. Of special interest have been **decentralized** decision and control problems that arise typically in connection with large-scale systems, and in particular with Naval C^3 systems. Included here have been our development of decentralized shortest path algorithms for networks, our examination of decentralized sequential detection and decision-making problems, and our studies of quantitative measures of interaction between system inputs and outputs (or between subsystems) and the implications of these for near disturbance localization and near decoupling and for estimation and control problems. Among the other problems we have examined are the use of the cutoff rate as a criterion for signal design in coded, direct-detection optical communication, a unified treatment of nonlinear smoothing problems with a variety of state and observation models, and the design of a *minimum-order* observer that estimates a prescribed multi-dimensional linear function of the system state.

We now review in turn our research in each of these areas. Since detailed developments are available in published papers or in drafts submitted, or to be submitted, for publication, we simply outline the results obtained and refer to the appropriate papers, which are included as appendices, for more complete and detailed expositions. The six major topic areas we discuss are:

Optimum Signal Design for Coded, Direct-Detection Optical Communication Systems.

Informationally-Decentralized Network Problems.

Decentralized Estimation and Control.

Nonlinear Smoothing Algorithms.

Minimum-Order Observers and State Estimators

Decentralized Sequential Detection and Decision Making.

II. ACCOMPLISHMENTS

Optimum Signal Design for Coded, Direct-Detection Optical Communication Systems.

This research was concerned with the coordinated design of the encoder, optical modulator and demodulator for a digital communication system employing an optical carrier and direct detection. The basis of our new approach was the reformulation of this signal design problem to use the cutoff rate as the performance measure instead of the usually-employed probability of error. We derived the cutoff rate for a digital communication system employing an optical carrier and direct detection, and we used this as the performance measure in studying the coordinated design of the optical modulator and demodulator. The choice of modulation that maximizes the cutoff rate was derived for various relationships between peak amplitude and average energy constraints on the transmitted optical signal. In particular, we showed that pulse position modulation is optimum when the average energy constraint is predominant, while Hadamard matrices can be used to define an optimum choice of modulation when the peak amplitude constraint predominates. We also addressed within this framework problems of efficient energy utilization, the choice of the dimensions of the input and output alphabets, and the effect of random detector gain.

The results of this extended research effort are contained in the journal paper:

> "Some Implications of the Cutoff-Rate Criterion for Coded, Direct Detection Optical Communication Systems", Donald L. Snyder and Ian B. Rhodes, *IEEE Transactions on Information Theory*, **IT-26**, No. 3, pp. 327-338, May 1980.

These same ideas and analysis were also extended to the situation where polarization modulation is employed in the optical modulator as well as temporal modulation, and completely analogous results were shown to hold. Specifically, for an input alphabet of dimension 4, the optimal modulation when average signal energy constraints predominate employs binary pulse-position and binary polarization modulation; such a modulation scheme has in fact been employed in gigabit-per-second satellite optical communication systems.

These results for polarization modulation were presented in the conference paper:

> "Quaternary Pulse Modulation is Optimal for Optical Communication at One Gigabit-per-Second", Donald L. Snyder and Ian B. Rhodes, National Communications Conference, Washington, D.C., November 27-29, 1979.

Informationally-Decentralized Network Problems.

This research effort was concerned with the development of informationally-decentralized shortest path algorithms that enable each node in a network to find its shortest distance to any other node using only local knowledge of the network topology and only local information transfer between adjacent nodes. We succeeded in devising several dynamic algorithms that accomplish these goals and can accommodate multiple changes in the network, including not only branch length increases and decreases but also topological changes such as the loss of branches or nodes and the introduction or reintroduction of branches or nodes. The ability of an algorithm to handle such changes while retaining its informationally- and topologically-decentralized character is essential in many practical applications, including especially those that arise in connection with C^3 systems. Furthermore, all of our algorithms operate asynchronously and converge to the optimum in finite time.

All of our algorithms can be considered to be modifications of or alternatives to our main dynamic algorithm which can accommodate all possible topological changes in the network, operates asynchronously, has localized information and communication requirements, and is guaranteed to converge in finite time. These modified or alternative algorithms retain these properties but differ in their complexity, in their methods for handling the various topological changes that can occur, and in the specific applications for which they are most appropriate. These variations result from differences between the several alternative mechanisms we devised for propagating the effects of any change throughout the network; because the algorithms are to operate asynchronously and with minimal information transfer between adjacent nodes only, awareness of the change and incorporation of its effects propagates in a nondeterministic way through the network. In particular, a means must be provided to ensure that each affected node learns of any change that has occurred and subsequently accepts and propagates further only distance information that can be guaranteed to fully incorporate the effects of the change. Several different changes may take place simultaneously, including branch failures that may disrupt the propagation of the effects of other changes. The introduction of new branches or nodes into a network in which the effects of other changes may still be propagating is an especially delicate matter that requires sophisticated control mechanisms if difficulties are to be avoided and convergence guaranteed.

These algorithms are reported in detail in the journal paper:

"Some Shortest Path Algorithms with Decentralized Information and Communication Requirements", Jeffrey M. Abram and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-27**, No. 3, pp. 570-582, June 1982.

Some of these results were also given in the conference presentation:

> "Some Informationally-Decentralized Network Algorithms", Jeffrey M. Abram and Ian B. Rhodes, Proceedings of the 1980 Joint Automatic Control Conference, San Francisco, California, August 13-15, 1980.

Decentralized Estimation and Control

This research has been motivated by our interest in decentralized estimation, decision and control problems. Such problems arise typically in connection with large systems, including especially C^3 systems. In other sections of this report we discuss our development of decentralized algorithms for determining the shortest paths in a network and our investigation of decentralized detection and decision-making problems. In this section we review our examination of the structural properties of linear systems and the implications of these in the analysis and design of decentralized estimation and control algorithms.

The underlying basis of this research is our development of quantitative measures of reachability and observability. This work was presented at *the 8th IFAC Congress*:

> "Some Quantitative Measures of Controllability and Observability and their Implications", Ian B. Rhodes, Proceedings of the Eighth Triennial World Congress of the International Federation of Automatic Control, Kyoto, Japan, August 24-28, 1981.

and an earlier version at the MIT/ONR Workshop on C^3 Problems:

> "Recent Results in Estimation Theory", Ian B. Rhodes, Third MIT/ONR Workshop on Distributed Information and Decision Systems Motivated by Command-Control-Communication Problems, Washington, D.C., May/June 1980.

It provides measures of reachabilty and unreachability, and of observability and unobservability, using Fenchel duality theory: the usual system-theoretic dualities are also preserved by these measures.

Within this framework we have quantified the degree of interaction or noninteraction between system input and output. The sum of the second-order modes of the system, interpreted as a measure of the overall strength of interaction between system inputs and outputs, arises naturally when the established theories of completely noninteracting system design (which are based on subspace inclusion relations) are generalized to design based approximate noninteraction using continuous-valued measures. In other words, this framework has provided the setting for generalizing the body of results concerning disturbance rejection and decoupling by providing a setting where questions of degree can be addressed and not just questions of kind.

In this work, the sum of the second-order modes is taken as the central cost functional in a theory of near-disturbance localization. In the case that exact disturbance localization is possible, we recover the results of geometric state-space theory. For minimum-phase systems satisfying the uniform rank condition, the results we obtain agree with those obtained with the "cheap control" approach to near non-interaction. In this case we obtain a control law of the same type as the solutions to the exact disturbance localization problem, which furthermore reduces to a deadbeat control on the "observable subspace". The optimal feedback gain is also expressible as the limit, as the weighting on the control energy goes to zero, of the solution to a linear-quadratic regulator problem. The optimal feedback gains are also shown to satisfy a first-order necessary condition.

These results have been outlined in the conference presentation:

> "Near Disturbance Localization Using Second-Order Modes", Joan M. Saniuk and Ian B. Rhodes, Proceedings of the 23rd Annual Allerton Conference on Control, Communications and Computing, University of Illinois, Oct. 1985.

A full length journal paper is in preparation. These studies have also led to the development of an ancillary result which is presented in:

> "A Matrix Inequality Associated with Bounds on Solutions of Algebraic Riccati and Lyapunov Equations", Joan M. Saniuk and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-32**, No. 8, Aug. 1987, pp. 739-740.

Nonlinear Smoothing Algorithms.

This research effort has produced a unified treatment of the smoothing problem for large classes of state and measurement processes. The state processes $\{x_t\}$ considered are of two types: (i) processes evolving in accordance with a diffusion equation of the Ito type, and (ii) continuous-time, finite-state Markov processes. The observation processes considered are also of two types: (i) the familiar observation model consisting of a nonlinear function of the state additively corrupted by white Gaussian noise, and (ii) a doubly-stochastic counting process, the rate of which is dependent on the state.

Within this unified framework, we have derived a number of alternative representations of the solution to the nonlinear smoothing problem. These include as special cases the numerous existing results for specific choices of state and observation processes, and they extend or generalize these results to all possible combinations cf state and measurement processes covered by this single, comprehensive formulation.

The alternative representations are:

(1) A symmetric representation of the solution to the smoothing problem in terms of the solutions to two filtering problems, one of which evolves forwards in time and the other backwards in time. The first of these is simply the conventional filter that estimates the present state in terms of past observations. The second estimates the present state in terms of future observations, and it is just like the first except that it operates in reverse time and requires models of the state and observation processes that are equivalent to the given models but evolve backwards in time. We provide the construction of such reverse-time models by appropriately generalizing and extending available results to encompass the full range of state and observation models that we consider.

(2) An asymmetric representation of the solution to the smoothing problem in terms of a conventional filter that runs forward in time and a likelihood ratio that evolves backwards in time. Again, the backwards evolution of the likelihood ratio is easily determined once equivalent reverse-time models for the state and observation processes are at hand.

(3) A representation that is not driven by tre observation process and that expresses the solution of the smoothing problem in terms of the solution to the forward filtering problem. This solution involves first processing the observations using a conventional forward filter, and then generating the smoothed estimates from these forward-filtered estimates from a calculation that evolves backwards in time but does not require direct use of the observations.

These results have been published in the journal article:

> "Smoothing Algorithms for Nonlinear Finite-Dimensional Systems", Brian D. O. Anderson and Ian B. Rhodes, *Stochastics*, Vol. 9, 1983, pp. 139 - 165.

and an abbreviated conference counterpart

> "Nonlinear Smoothing Algorithms", Brian. D. O. Anderson and Ian B. Rhodes, 1983 IEEE International Symposium on Information Theory, Les Arcs, France, June 1983.

Minimum-Order Observers and State Estimators.

This research has addressed a long-standing problem, that of designing a minimum-order observer that asymptotically reconstructs a specified multi-dimensional function of the state of the observed system. Our initial motivation for examining this problem lay in the application of such observers in the design and performance evaluation of decentralized controllers for non-interacting and nearly non-interacting systems. What resulted has been a contribution to the fundamental problem of minimum-order observer design. We have developed algorithms for the design of both stable minimum-order observers and minimum-order observers with arbitrary dynamics, as well as a collection of analytical results concerning such issues as uniqueness and bounds on the dimension of the observer. Interestingly, the algorithms we have developed for the standard "centralized" minimum-order design problem make use of results from decentralized control theory.

The observer design problem is as follows:

For the constant linear system

$$\dot{x}(t) = Ax(t); \qquad x(t) \in R^n$$

$$y(t) = Cx(t); \qquad y(t) \in R^m$$

design an observer

$$\dot{z}(t) = Fz(t) + Gy(t); \qquad z(t) \in R^r$$

that asymptotically reconstructs the prescribed q-dimensional function $Kx(t)$ of the state in the sense that

$$Mz(t) + Ny(t) \longrightarrow Kx(t) \; as \; t \longrightarrow \infty$$

The entities to be chosen are the dimension r of the observer and the matrices F, G, M and N. The objective is for r to minimum consistent with either (i) F is stable, or (ii) the cigenvalues of F are assignable arbitrarily. In the former case we require simply that the observer be stable, while in the latter we incorporate complete flexibility in the choice of the observer dynamics.

These two observer design problems can be expressed in the framework of geometric linear system theory: the first problem becomes one of finding a minimal (A',C')-invariant subspace τ that, together with the range of C', contains the range of K' and, for some L such that $(A+LC)'\tau \subset \tau$, $(A+LC)'|_\tau$ is stable. In the language of geometric linear system theory, τ is a minimal stable *cover* of the range of K'. The second problem becomes one of finding a minimal (A',C')–*controllability* subspace τ that, together with the range of C', contains the range of K'. In this case, we seek a minimum cover that is also a controllability subspace. Thus, results for minimum-order observer design have counterparts in the geometric theory of linear systems, and *vice-versa.*

The algorithm we have developed for constructing a minimal stable observer (cover) makes use of standard results for $(n-m)$-dimensional reduced-order observers. The idea underlying our approach is as follows: we begin by temporarily augmenting the available output y with certain artificial outputs, thus creating an augmented output y_a with dimension m_a

that is used to drive a standard $(n-m_a)$-dimensional reduced-order observer. This reduced-order observer is then modified, keeping its dimension fixed, by removing the artificial outputs that were temporarily added until only the available output y remains driving it. The removal of each artificial output results in constraints on the observer dynamics that reflect the requirement that the prescribed linear function $Kx(t)$ be asymptotically reconstructible from y and the state of the modified observer. The result is a characterization of all covers (or, equivalently, of all not necessarily stable "observers" of dimension $n-m_a$) that meet the requirements on estimating Kx. If there is a stable observer amongst these then no more need be done. If not, the number of artificially-added outputs is reduced by one, thus increasing the dimension $(n-m_a)$ of the observer by one, and the process repeated. The observer dimension that is used initially in this iterative process is provided by a new lower bound that we have derived on the dimension of the minimum stable observer, while a known upper bound limits the number of iterations that are needed before a stable observer is found. We have also obtained new results on the uniqueness of the minimal cover and of the F matrix (modulo a similarity transformation) in the minimal observer.

The above characterization of all covers of a given dimension that meet the requirements on estimating Kx also provides the basis for our solution of the second problem: we seek from among these one that is also a controllability subspace. We do this by reformulating the problem in such a way that we can bring to bear a result of Corfmat and Morse that pertains to the decentralized control of linear systems. In fact, the checking of the Corfmat-Morse conditions in this case can be simplified considerably by exploiting the particular structure that arises here. Again, the overall process is an iterative one: if no controllability subspace can be found from among the covers of a given dimension, then none exists and the dimension is increased by one and the process is repeated. We have derived lower and upper bounds on the dimension of the smallest controllability subspace that covers the range of K'; this provides an initial observer dimension for this iterative process and an upper limit on the number of iterations before a solution is found.

The results on the uniqueness of the minimal cover are available in a conference proceedings:

> "Uniqueness of the Minimal Dynamic Cover and the Associated Solution to Sylvester's Equation", Amr F. Assal and Ian B. Rhodes, Proceedings of the 24th Annual Allerton Conference on Control, Communications and Computing, University of Illinois, Oct. 1986.

Two journal papers describing the results of the major part of this research effort are currently in preparation.

Decentralized Sequential Detection and Decision Making.

Problems of decision making in the face of statistical uncertainties have long been of interest to decision theorists in various disciplines. Detection theory is one area of decision making that has received much attention, particularly in surveillance systems. The well-known theory of classical detection has been successfully applied to single-sensor detection problems and to multiple-sensor problems when all collected data is transmitted to a central site for processing. In those cases where it is impractical or infeasible for all the raw data collected at the local sites to be transmitted to a central processing center, a decentralized (or distributed) detection problem results in which each of the local sites is asked to perform some preprocessing of its own raw data before communicating with the fusion center. Although suboptimal compared with centralized detection because of the loss of information in local processors, decentralized detection is in many cases a more realistic formulation in practical applications than its centralized counterpart. For example, in the face of capacity-constrained channels, local processing could substantially decrease the communication bandwidth to the central site, thereby speeding up the process and reducing the communication costs. Also, decentralization may be the only natural way to model the problem in situations where multiple detectors of various types are located at dispersed geographical sites. Furthermore, decentralized detection may be imposed in situations where, due to enormous amounts of available raw data, centralized processing of the information is not feasible. Other issues include potential system reliability and integrity in the face of failures.

There are a variety of ways in which one can pose a decentralized detect'on problem. Each formulation has elements that capture one or more of the following three features of decentralized detection and decision-making problems:

Hierarchical Structure: Local observations are processed and the result sent to a fusion center where a final or "global" decision is made. There may or may not be any direct communication between the local observers.

Multi-Stage or Sequential Structure: Additional measurements can be taken by the local observers. Each local observer might wait until he elects to stop taking measurements before sending information to the fusion center, or information might be sent after each observation and the fusion center given the flexibility to determine when to declare a decision.

Information Rate or Bandwidth Reduction: The preprocessing performed by each local observer should result in a significant reduction in the amount of data sent to the fusion center. An extreme example is for each local site to make a "local decision" and to communicate this to the fusion center.

We have analyzed a broad class of multi-stage decentralized detection and decision-making problems that captures for the first time all three of the above elements. The most general of these is a multi-stage, multi-detector decentralized binary hypothesis-testing problem in which each detector, after obtaining each of his observations, sends a binary decision (0 or 1) to the fusion center or "supervisor", who is given the option of declaring a final decision (as to which of two possible hypotheses is true) at any stage depending on the local decisions that have been received. There is a cost associated with each combination of the true hypothesis and the final decision. In addition, a cost is incurred s delayed until the next time instant. There is no communication among the local detectors.

We have shown that, under appropriate independence assumptions that are in many practical applications quite reasonable, the optimal local strategies are governed by threshold tests on the likelihood ratio (i.e., likelihood-ratio tests), where the thresholds are precomputable off-line. This is an important result since it makes tractable an otherwise intractable problem. Indeed, the most general problem is known to be NP-complete, and the assumption that the local observations are conditionally independent given the hypotheses is crucial in providing the simplification of the optimal strategies to likelihood ratio tests.

These results are presented in two publications. The first,

> "Decentralized Sequential Detection", H. R. Hashemi and Ian B. Rhodes, *IEEE Transactions on Information Theory*, to appear.

develops the results for the two-detector, two-stage problem, along with some of the important properties of the optimal strategies that suggest a significant reduction in the computations required to determine the solution. The second,

> "Decentralized Dynamic Decisicn Making", H. R. Hashemi and Ian B. Rhodes, Proceedings of the 1987 IEEE Conference on Decision and Control, Los Angeles, 1987, pp. 1836-1841 (Invited Paper).

contains the results for the general multi-detector, multi-stage problem.

III. THESES SUPPORTED

The following doctoral dissertations have been supported or partially supported under this Contract.

"Shortest Path Algorithms with Decentralized Information and Communication Requirements", Jeffrey M. Abram, D.Sc. Dissertation, Washington University, St. Louis, May 1981.

"Decentralized Sequential Detection", Hamid R. Hashemipour, Ph.D. Dissertation, University of California, Santa Barbara, December 1985.

"Issues in the Design of Reduced-Order Observers", Amr F. Assal, Ph.D. Dissertation, University of California, Santa Barbara, March 1986.

"Contributions to a Theory of Nearly Non-Interacting Systems", Joan M. Saniuk, Ph.D. Dissertation, University of California, Santa Barbara, October 1986.

IV. PUBLICATIONS UNDER CONTRACT N00014-79-C-0549

"Quaternary Pulse Modulation is Optimal for Optical Communication at One Gigabit-per-Second", Donald L. Snyder and Ian B. Rhodes, National Communications Conference, Washington, D.C., November 27-29, 1979.

"Some Implications of the Cutoff-Rate Criterion for Coded, Direct Detection Optical Communication Systems", Donald L. Snyder and Ian B. Rhodes, *IEEE Transactions on Information Theory*, **IT-26**, No. 3, pp. 327-338, May 1980.

"Recent Results in Estimation Theory", Ian B. Rhodes, Third MIT/ONR Workshop on Distributed Information and Decision Systems Motivated by Command-Control-Communication Problems, Washington, D.C., May/June 1980.

"Some Informationally-Decentralized Network Algorithms", Jeffrey M. Abram and Ian B. Rhodes, Proceedings of the 1980 Joint Automatic Control Conference, San Francisco, California, August 13-15, 1980.

"Some Quantitative Measures of Controllability and Observability and their Implications", Ian B. Rhodes, Proceedings of the Eighth Triennial World Congress of the International Federation of Automatic Control, Kyoto, Japan, August 24-28, 1981.

"Some Shortest Path Algorithms with Decentralized Information and Communication Requirements", Jeffrey M. Abram and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-27**, No. 3, pp. 570-582, June 1982.

"Nonlinear Smoothing Algorithms", Brian D. O. Anderson and Ian B. Rhodes, 1982 IEEE International Symposium on Information Theory, Les Arcs, France, June 21-25, 1982.

"Smoothing Algorithms for Finite-Dimensional Systems", Brian D. O. Anderson and Ian B. Rhodes, *Stochastics*, Vol. 9, pp. 139-165, 1983.

"Near Disturbance Localization Using Second-Order Modes", Joan M. Saniuk and Ian B. Rhodes, Proceedings of the 23rd Annual Allerton Conference on Control, Communications and Computing, University of Illinois, Oct. 1985.

"Uniqueness of the Minimal Dynamic Cover and the Associated Solution to Sylvester's Equation", Amr F. Assal and Ian B. Rhodes, Proceedings of the 24th Annual Allerton Conference on Control, Communications and Computing, University of Illinois, Oct. 1986.

"A Matrix Inequality Associated with Bounds on Solutions of Algebraic Riccati and Lyapunov Equations", Joan M. Saniuk and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-32**, No. 8, Aug. 1987, pp. 739-740.

"Decentralized Sequential Detection", H. R. Hashemi and Ian B. Rhodes, *IEEE Transactions on Information Theory*, to appear.

"Decentralized Dynamic Decision Making", H. R. Hashemi and Ian B. Rhodes, Proceedings of the 1987 IEEE Conference on Decision and Control, Los Angeles, 1987, pp. 1836 - 1841 (Invited Paper).

V. APPENDIX

Copies of Publications

Reprint of Abstract:

"Quaternary Pulse Modulation is Optimal for Optical Communication at One Gigabit-per-Second", Donald L. Snyder and Ian B. Rhodes, National Telecommunications Conference, Washington, D.C., November 27-29, 1979.

Quaternary Pulse Modulation Is Optimal for Optical Communication at One Gigabit per Second*

Donald L. Snyder**

Ian B. Rhodes***

SUMMARY

The received data in an optical communication system employing ideal direct detection is modeled by a Poisson random point process, which accounts for both the quantum nature of light and the optical-to-electrical energy conversion process in the detector. Using this model, we have evaluated the computational cutoff rate of the discrete channel formed by an encoder, optical modulator-channel-demodulator, and decoder. We find that quaternary pulse modulation, a 4-ary scheme employing binary pulse-position and binary polarization modulation, maximizes this cutoff rate when the optical field produced by the modulator is subject to peak-value and average-energy constraints. Such a modulation format has been adopted in a one gigabit per second satellite optical communication system under development for the Air Force.

* This work was supported by the National Science Foundation under Grants ENG76-11565 and ENG 78-19885 and by the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources, and the Office of Naval Research under Contract N00014-76-C-0667.

** Department of Electrical Engineering and Biomedical Computer Laboratory, Washington University, St. Louis, Mo. 63130.

*** Department of Systems Science and Mathematics, Washington University, St. Louis, Mo. 63130.

Reprint of Paper:

"Some Implications of the Cutoff-Rate Criterion for Coded Direct-Detection Optical Communication Systems ", Donald L. Snyder and Ian B. Rhodes, *IEEE Transactions on Information Theory*, **IT-26**, No. 3, pp. 327-338, May 1980.

Some Implications of the Cutoff-Rate Criterion for Coded Direct-Detection Optical Communication Systems

DONALD L. SNYDER, SENIOR MEMBER, IEEE, AND IAN B. RHODES, MEMBER, IEEE

Abstract—The cutoff rate is derived for a digital communication system employing an optical carrier and direct detection. The coordinated design of the encoder, optical modulator, and demodulator is then studied using the cutoff rate as a performance measure rather than the more commonly employed error probability. Modulator design is studied when transmitted optical signals are subject simultaneously to average-energy and peak-value constraints. Pulse-position modulation is shown to maximize the cutoff rate when the average-energy constraint predominates, and the best signals when the peak-value constraint predominates are identified in terms of Hadamard matrices. A time-sharing of these signals maximizes the cutoff rate when neither constraint dominates the other. Problems of efficient energy utilization, choice of input and output alphabet dimension, and the effect of random detector gain are addressed.

I. Introduction

OUR CONCERN in this paper is with digital communication systems that employ coherent light as a carrier and direct detection as the means to convert the received optical field into an electrical signal for subsequent processing. Communication systems of this type are discussed widely in the literature (see [1]–[5] and references therein) and are of increasing importance in applications. The optical portion of the overall system consists of the optical modulator, optical channel, and optical detector shown schematically in the basic information-theoretic model of the optical digital communication system of Fig. 1. Here, $E(t,\vec{r})$ represents the temporally and spatially dependent complex envelope of the optical field, and $N(t)$ represents the counting process associated with the output of an ideal direct-detection device. This counting process is assumed to be an inhomogeneous Poisson process with rate function $\lambda(t)=s(t)+\lambda_0$, where λ_0 represents the contribution to the total count rate due to dark current in the detector. Also, λ_0 can account for background radiation when this is characterized by many weak modal-components [2], [3]. The assumption that $N(t)$ is a Poisson process is met to a close approximation on the free-space channel for coherent sources [3]. On our model the signal count rate $s(t)$ is related to $E(t,\vec{r})$ according to

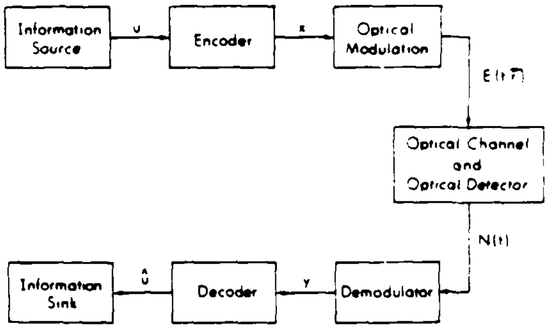


Fig. 1. Optical digital communication system.

$$s(t)=(\eta/h\nu)\int_A|E(t,\vec{r})|^2\,d\vec{r}, \tag{1}$$

where η is the quantum efficiency of the detector, h is Planck's constant, ν is the optical-carrier frequency, and A is the active surface of the detector; it is evident that $s(t)$ is nonnegative, which, of course, it must be as a rate function.

We shall suppose that a code letter x in Fig. 1 is drawn once each T seconds from a $q$ary alphabet $\mathcal{X}=\{X_1,X_2,\cdots,X_q\}$. We further suppose that each demodulator-output letter y is drawn from a q'ary alphabet $\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_{q'}\}$, where in general $q'>q$. Initially, we investigate "infinitely soft" decisions for which $q'=\infty$; then we study the penalty for choosing a smaller value of q'. The decoder output letters $\hat{u}$ supplied to the sink are reproductions of the encoder input letters u supplied by the source; these are presumed to be drawn from a binary alphabet $\mathcal{U}=\{0,1\}$. The rate of the coding system in terms of the number of source digits for each channel letter will be denoted by R bits per channel use. This means that $R=R_tT$ if the source generates R_t bits per second.

The combination of the optical modulator, optical channel, optical detector, and demodulator forms a discrete channel with a $q$ary input alphabet $\mathcal{X}$ and q'ary

Manuscript received April 7, 1978; revised April 9, 1979. This work was supported in part by the National Science Foundation under Grant ENG 76-11565, in part by the National Institutes of Health, Division of Research Resources, under Research Grant RR 00396, and in part by the Office of Naval Research under Contract N00014-79-C-0549.

D. L. Snyder is with the Department of Electrical Engineering and Biomedical Computer Laboratory, Washington University, P.O. Box 1127, St. Louis, MO 63130.

I. B. Rhodes was with the Department of Electrical Engineering, Washington University, St. Louis, MO. He is now with the Department of Electrical Engineering, University of California, Santa Barbara, CA 93106.

0018-9448/80/0500-0327$00.75 ©1980 IEEE

output alphabet $\mathcal{Y}$. By virtue of the independent-increments property of the Poisson process and the constancy of λ_0, this is a "constant discrete memoryless channel" in the sense that the conditional probability the channel output sequence is $b_1 b_2 \cdots b_n$, where each b_i is in $\mathcal{Y}$, given that the input sequence is $a_1 a_2 \cdots a_n$, where each a_i is in $\mathcal{X}$, factors into the n-fold product of the per-letter transition probabilities according to

$$\Pr(y_1=b_1, y_2=b_2, \cdots, y_n=b_n \mid x_1=a_1, x_2=a_2, \cdots, x_n=a_n) = \prod_{i=1}^{n} \Pr[\,y_i=b_i \mid x_i=a_i\,]. \qquad (2)$$

Furthermore, the per-letter transition probabilities are the same for any T second use of the channel. Thus, if $p_{y|x}(Y|X)$ denotes the per letter transition probability, the right side of (2) is $\prod_{i=1}^{n} p_{y|x}(b_i|a_i)$. The design of the modulator and demodulator, of course, affects $p_{y|x}(Y|X)$. We shall study the design which makes $p_{y|x}(Y|X)$ most favorable for a given optical channel and detector. The coordination of this design with that of the encoder will also be studied.

A quantity that reflects the influence of $p_{y|x}(Y|X)$ on the quality of a constant discrete memoryless channel is the cutoff rate R_0 defined by

$$\begin{aligned} R_0 &= -\log_2\left\{ \min_{Q} \sum_{Y\in\mathcal{Y}} \left[\sum_{X\in\mathcal{X}} (p_{y|x}(Y|X))^{1/2} Q(X) \right]^2 \right\} \\ &= -\log_2\left\{ \min_{Q} \sum_{i=1}^{q} \sum_{j=1}^{q} Q(X_i)Q(X_j) \right. \\ &\qquad \left. \cdot \sum_{k=1}^{q} (p_{v|x}(Y_k|X_i) p_{y|x}(Y_k|X_j))^{1/2} \right\}, \end{aligned} \qquad (3)$$

where Q is a probability mass function on $\mathcal{X}$. Wozencraft and Kennedy [6], in 1966, were first to argue in favor of the cutoff rate as a criterion for design because it is the upper limit of code rates R for which the average decoding computation per source digit is finite when sequential decoding is used. Wozencraft and Kennedy also showed that there is a block code of rate R and codeword length N such that the probability of error $\Pr(e)$ in decoding a sourceword of length $K=NR$ is bounded according to

$$\Pr(e) \le 2^{-N(R_0-R)}, \qquad \text{if } R<R_0. \qquad (4)$$

Thus, for block codes, the single number R_0 provides a measure of both a range of rates R for which reliable communication is possible as well as the coding complexity, as reflected by N, required to guarantee a specified block error probability. More recently, Viterbi [7] has shown for convolutional coding and maximum-likelihood sequence decoding on the constant discrete memoryless channel that the error probability is upper bounded according to

$$\Pr(e) \le C_R L 2^{-NR_0} \qquad \text{if } R<R_0, \qquad (5)$$

where N is the constraint length of the convolutional code, R is the code rate, L is the total number of source letters encoded, and C_R is a weakly dependent function of R and not a function of L and N. Thus, as with block codes, the single number R_0 provides a measure of both reliable rates and code complexity. Massey [8], [9] made these observations first and has used them to make an eloquent and persuasive argument for adopting R_0 as a modulator-demodulator design parameter in place of the more commonly used error probability. In what follows we shall investigate some of the implications of attempting to maximize this parameter for modulator–demodulator design for direct-detection optical communication systems.

II. R_0 for Infinitely Fine Quantization

In practice, the demodulator of Fig. 1 must quantize the point process observed on $[0,T]$ in some fashion to produce one of the q' output letters in $\mathcal{Y}$. This might be accomplished, for example, by counting points in subintervals of $[0,T]$, disregarding their times of occurrence within these subintervals, and then comparing the subinterval counts to prescribed thresholds. Regardless of what form of quantization is adopted, the finer it is, the larger will be the cutoff rate R_0 of the resulting constant discrete memoryless channel. Thus we consider first the limiting situation of infinitely fine quantization, for which $q'=\infty$ and $R_0 \equiv R_{0,\infty}$ is not degraded by quantization. Then we consider the effect of finite quantization.

For a Poisson process with rate $\lambda(t)$, the probability of observing n points during $[0,T]$ in n disjoint intervals $[t_1, t_1+\Delta t_1), [t_2, t_2+\Delta t_2), \cdots, [t_n, t_n+\Delta t_n)$ is approximated to $o(\max_i \Delta t_i)$ by

$$\left(\prod_{l=1}^{n} \lambda(t_l) \right) \exp\left(-\int_0^T \lambda(t)\,dt \right) \Delta t_1 \Delta t_2 \cdots \Delta t_n$$

for $n \ge 1$ and by

$$\exp\left(-\int_0^T \lambda(t)\,dt \right)$$

for $n=0$. Consequently, for infinitely fine quantization, the summation, call it $f(i,j)$, over k in (3) becomes

$$\begin{aligned} f(i,j) &= \exp\left(-\frac{1}{2}\int_0^T (\lambda_i(t)+\lambda_j(t))\,dt \right) \\ &\quad \cdot \left[1 + \sum_{n=1}^{\infty} \int\!\!\int \cdots \int \prod_{l=1}^{n} (\lambda_i(t_l)\lambda_j(t_l))^{1/2} dt_1\,dt_2 \cdots dt_n \right], \end{aligned}$$

where $\lambda_i(t)$ and $\lambda_j(t)$ are the detection rates for code letters X_i and X_j, respectively, and the integration is over the region $0 \le t_1 \le t_2 \le \cdots \le t_n \le T$. By extending this range of integration to $0 \le t_i \le T$ for $i=1,2,\cdots,n$, and dividing by $n!$ to compensate for this extension, we obtain

$$f(i,j) = \exp\left(-\frac{1}{2}\int_0^T (g_i(t)-g_j(t))^2\,dt \right),$$

where we define $g_i(t)=\lambda_i^{1/2}(t)$ and $g_j(t)=\lambda_j^{1/2}(t)$. Thus

$$R_{0,\infty} = -\log_2\left\{ \min_{Q} \sum_{i=1}^{q} \sum_{j=1}^{q} Q(X_i)Q(X_j) \cdot \exp\left(-\frac{1}{2}\int_0^T (g_i(t)-g_j(t))^2\,dt \right) \right\}. \qquad (6)$$

This expression is identical to that obtained by Massey [8, eq. (4)] if the signal $g_i(t)$ were to be observed in an additive white Gaussian noise of unit intensity when X_i is the code letter into the modulator. It is with this expression that Massey established for the first time the $R_{0,\infty}$-optimality under an average energy constraint of a simplex signal set for the additive white Gaussian noise channel. However, the additional constraint $g_i(t) > \lambda_0^{1/2} > 0$ obtains here, so Massey's argument does not hold for direct-detection optical communication systems and must be modified. This is accomplished as follows.

By defining

$$S = \sum_{i=1}^{q} Q^2(X_i)$$

and by using Jensen's inequality, Massey [8] shows from (6) that

$$R_{0,\infty} < -\log_2\left\{ \min_Q \left[S + (1-S)\exp\left(-\frac{1}{2(1-S)} \cdot \sum_{i=1}^{q}\sum_{j=1}^{q} Q(X_i)Q(X_j)\int_0^T (g_i(t) - g_j(t))^2 dt \right) \right] \right\}, \quad (7)$$

with equality holding if and only if the quantity

$$d_{ij}^2 \triangleq \int_0^T (g_i(t) - g_j(t))^2 dt \quad (8)$$

is the same whenever $i \neq j$. It is evident from (7) that $R_{0,\infty}$ *is a monotonically increasing function of d_{ij} for $i \neq j$.* Thus, if d denotes the maximum of the d_{ij} for $i \neq j$,

$$R_{0,\infty} < -\log_2\left\{ \min_Q \left[S + (1-S)\exp\left(-\frac{1}{2}d^2 \right) \right] \right\}. \quad (9)$$

Furthermore, it is easily verified that the minimizing code letter distribution in (9) is the uniform distribution $Q(X_i) = 1/q$ for $i = 1, 2, \cdots, q$. As $S = 1/q$ for this distribution, (9) becomes

$$R_{0,\infty} < \log_2 q - \log_2\left[1 + (q-1)\exp\left(-\frac{1}{2}d^2 \right) \right], \quad (10)$$

with equality holding if and only if $d_{ij} = d$ whenever $i \neq j$.

III. Modulator Design Based on $R_{0,\infty}$

An optical modulator designed to produce a signal set $\mathcal{S} = \{E_1(t,\vec{r}), E_2(t,\vec{r}), \cdots, E_q(t,\vec{r})\}$ such that $d_{ij} = d$ for $i \neq j$ and such that d is as large as possible produces the best overall performance for the digital optical communication system as measured by $R_{0,\infty}$. Thus we are motivated to *examine the maximization of d subject to suitable constraints on signals in $\mathcal{S}$.* Associated with each signal set $\mathcal{S}$ is a derived signal set $\mathcal{G} = \{g_1(t), g_2(t), \cdots, g_q(t)\}$ in which $g_i(t) = \lambda_i^{1/2}(t)$, where

$$\lambda_i(t) = (\eta/h\nu)\int_A |E_i(t,\vec{r})|^2 d\vec{r} + \lambda_0. \quad (11)$$

Note that signals in $\mathcal{G}$ satisfy $g_i(t) > g_{\min} = \lambda_0^{1/2}$.

This maximization problem is examined subject to additional constraints on the average energy and the peak amplitude of signals in the transmitted signal set $\mathcal{S}$. We assume that the average energy $\bar{E}$ of signals in $\mathcal{S}$, defined by

$$\bar{E} = \frac{1}{q}\sum_{i=1}^{q}\int_0^T\int_A |E_i(t,\vec{r})|^2 d\vec{r}\,dt. \quad (12)$$

must satisfy

$$\bar{E} < \bar{E}_{\max}, \quad (13)$$

where $\bar{E}_{\max}$ is a prespecified maximum allowable average energy. Then the average energy $\bar{E}_g$ for signals in the derived signal set $\mathcal{G}$, defined by

$$\bar{E}_g = \frac{1}{q}\sum_{i=1}^{q}\int_0^T g_i^2(t)\,dt, \quad (14)$$

satisfies

$$\bar{E}_g - \bar{n} = \bar{s} < \bar{s}_{\max}, \quad (15)$$

where $\bar{s} = \eta\bar{E}/h\nu$ and $\bar{n} = g_{\min}^2 T = \lambda_0 T$ are the average number of signal counts and noise counts, respectively, per channel use, and where $\bar{s}_{\max} = \eta\bar{E}_{\max}/h\nu$. We assume, *further, that the amplitude $|E_i(t,\vec{r})|$ of each signal in the* transmitted-signal set $\mathcal{S}$ cannot exceed a prespecified maximum value $P_{\max}$; that is,

$$|E_i(t,\vec{r})| < P_{\max} \quad (16)$$

for $i = 1, 2, \cdots, q$, $0 < t < T$, and for all locations $\vec{r}$ in the active surface of the detector. Then each signal in the *derived signal set $\mathcal{G}$ satisfies*

$$g_{\min} < g_i(t) < g_{\max}, \quad (17)$$

where $g_{\min} = \lambda_0^{1/2}$ and $g_{\max} = [(\eta A/h\nu)P_{\max}^2 + \lambda_0]^{1/2}$.

For modulator design, we thus have the following optimization problem. Select signals in $\mathcal{G}$ to maximize

$$d^2 = [q(q-1)]^{-1}\sum_{i=1}^{q}\sum_{j=1}^{q}\int_0^T [g_i(t) - g_j(t)]^2 dt \quad (18)$$

subject to the following constraints.

i) *Equidistance constraint:* The quantities d_{ij} in (8) should be the same whenever $i \neq j$.

ii) *Average-energy constraint:* Equation (15) should be satisfied.

iii) *Peak-amplitude constraint:* Equation (17) should be satisfied.

To simplify the development we temporarily neglect the equidistance constraint in formulating and solving this optimization problem. It will be evident subsequently that among the solutions to the relaxed problem are ones *satisfying the equidistance constraint, and these are then* solutions to the fully constrained problem.

We find for $q' = \infty$ that the best choice of modulator design depends on the particular values of q, T, $g_{\min}$, $g_{\max}$, and $\bar{s}_{\max}$, but whatever values these parameters may be, there are only three categories of best design. These are determined by the conditions

$$\bar{s}_{\max} \in \left[0, \frac{1}{q}(g_{\max}^2 - g_{\min}^2)T \right], \quad (19a)$$

$$\bar{s}_{\max} \in \begin{cases} \left(\frac{1}{q}(g^2_{\max} - g^2_{\min}) T, \frac{1}{2}(g^2_{\max} - g^{\cdot\cdot}_{\min}) T \right), & q \text{ even}, \\ \left(\frac{1}{q}(g^2_{\max} - g^2_{\min}) T, \frac{q-1}{2q}(g^2_{\max} - g^2_{\min}) T \right), \\ q \text{ odd}, \end{cases} \tag{19b}$$

$$\bar{s}_{\max} \in \begin{cases} \left[\frac{1}{2}(g^2_{\max} - g^2_{\min}) T, \infty \right), & q \text{ even}, \\ \left[\frac{q-1}{2q}(g^2_{\max} - g^2_{\min}) T, \infty \right), & q \text{ odd}. \end{cases} \tag{19c}$$

We say that the "average-energy constraint predominates" when (19a) holds, the "peak-amplitude constraint predominates" when (19c) holds, and that "neither constraint predominates" when (1'9b) holds.

Average-Energy Constraint Predominates

We first give an upper bound on d^2 that holds regardless of which, if any, constraint predominates. Then we identify a modulator design that achieves this upper bound when the average-energy constraint predominates.

Suppressing the common argument t of all entities, we have

$$(g_i - g_j)^2 = (g_i - g_{\min})^2 - 2(g_i - g_{\min})(g_j - g_{\min}) + (g_j - g_{\min})^2$$
$$< (g_i - g_{\min})^2 + (g_j - g_{\min})^2,$$

the inequality holding because $(g_i - g_{\min}) > 0$ for all $i \in \{1, 2, \cdots, q\}$. Furthermore, for $i \neq j$, equality holds if and only if at most one of g_i and g_j is strictly greater than $g_{\min}$. Summing over $i \neq j$, then over $j \in \{1, 2, \cdots, q\}$, integrating over $[0, T]$ and dividing both sides by $q(q-1)$ yields

$$\frac{1}{q(q-1)} \int_0^T \sum_{i=1}^q \sum_{j=1}^q [g_i(t) - g_j(t)]^2 dt$$
$$< \frac{2}{q} \int_0^T \sum_{i=1}^q [g_i(t) - g_{\min}]^2 dt, \tag{20}$$

with equality holding if and only if, for almost all t, $g_i(t) > g_{\min}$ for at most one value of i in $\{1, 2, \cdots, q\}$. Now for any $i \in \{1, 2, \cdots, q\}$ and any $t \in [0, T]$,

$$[g_i(t) - g_{\min}][g_{\max} - g_i(t)] > 0,$$

which yields

$$g_i^2(t) - g_i(t)[g_{\max} + g_{\min}] < - g_{\max} g_{\min}, \tag{21}$$

with equality if and only if $g_i(t) = g_{\min}$ or $g_i(t) = g_{\max}$. We then have

$$[g_i(t) - g_{\min}]^2$$
$$= \frac{2g_{\min}}{g_{\max} + g_{\min}} [g_i^2(t) - (g_{\max} + g_{\min}) g_i(t)]$$
$$+ \left[1 - \frac{2g_{\min}}{g_{\max} + g_{\min}} \right] g_i^2(t) + g^2_{\min}$$
$$< \frac{2g_{\min}}{g_{\max} + g_{\min}} (- g_{\max} g_{\min}) + \left[\frac{g_{\max} - g_{\min}}{g_{\max} + g_{\min}} \right] g_i^2(t) + g^2_{\min}$$
$$= \left[\frac{g_{\max} - g_{\min}}{g_{\max} + g_{\min}} \right] [g_i^2(t) - g^2_{\min}], \tag{22}$$

where the inequality follows by virtue of (21), and equality holds if and only if $g_i(t) = g_{\max}$ or $g_i(t) = g_{\min}$. Finally, substituting (22) into (20), using the average-energy constraint (15), and noting the conditions for equality yields the following lemma.

Lemma 1: Let $\bar{s}_{\max}$ be the maximum average signal counts per channel use, and suppose $g_{\min} < g_i(t) < g_{\max}$ for $t \in [0, T]$ and $i \in 0, 2, \cdots, q$. Then

$$d^2 < 2 \left[\frac{g_{\max} - g_{\min}}{g_{\max} + g_{\min}} \right] \bar{s}_{\max}. \tag{23}$$

Furthermore, equality holds if and only if both a) at any time $t \in [0, T]$, all signals in $\mathcal{G}$ take on value $g_{\min}$ except at most one which takes on value $g_{\max}$; and b)

$$\frac{1}{q} \sum_{i=1}^q \int_0^T g_i^2(t) dt - g^2_{\min} T = \bar{s}_{\max}. \tag{24}$$

The condition a) for equality is simply a combination of the conditions for equality of (20) and (22), while condition b) is the condition for equality in (15).

A signal set that is equidistant and achieves the upper bound in Lemma 1 with equality, and which therefore maximizes the cutoff rate $R_{0,\infty}$ when the average-energy constraint predominates, is characterized in the following lemma.

Lemma 2: If $\bar{s}_{\max}$ satisfies (19a), equality is achieved in (23) by the equidistant pulse-position modulation (PPM) signal set

$$g_i^*(t) = \begin{cases} g_{\max}, & (i-1)T/q < t < (i - 1 + \epsilon)T/q, \\ g_{\min}, & \text{otherwise for } 0 < t < T, \end{cases} \tag{25}$$

where

$$\epsilon = \frac{q\bar{s}_{\max}}{T(g^2_{\max} - g^2_{\min})}. \tag{26}$$

To establish Lemma 2, note that the PPM signal set is clearly equidistant and that condition (19a) is equivalent to $\epsilon < 1$ so that condition a) in Lemma 1 is satisfied. Moreover the average number of signal counts per channel use for the PPM signal set is given by

$$\bar{s}^* = \frac{1}{q} \sum_{i=1}^q \int_0^T g_i^{*2}(t) dt - g^2_{\min} T = \frac{\epsilon T}{q} (g^2_{\max} - g^2_{\min}).$$

Hence, from (26), $\bar{s}^* = \bar{s}_{\max}$, so condition b) of Lemma 1 is also satisfied. Consequently, by Lemma 1, d^2 for this signal set equals the upper bound in (23). It is straightforward to verify by direct calculation for this PPM signal set and $\epsilon < 1$ that d^2 equals the upper bound. Lemma 2 follows, and we conclude that the equidistant PPM signal set (25) maximizes $R_{0,\infty}$ when the average-energy constraint predominates.

Lemma 2 can be strengthened by noting that the equidistant PPM signal set (25) is the unique signal set that achieves equality in (23) modulo shifting or splitting of pulses while keeping them nonoverlapping and keeping the total "on-time" of any g_i equal to $\epsilon T/q$. This is because condition a) of Lemma 1 is satisfied if and only if the pulses are nonoverlapping and because $\epsilon < 1$ is chosen precisely to use up all the available energy, as required by condition b) of Lemma 1.

Peak-Amplitude Constraint Predominates

By this we mean that the energy constraint (15) is not a limiting consideration. We therefore neglect it, as well as the equidistance constraint, and consider the problem of maximizing d^2 in (18) subject only to (17). The average energy required by the signals that solve this problem will then provide conditions for dominance of the peak-amplitude constraint, and among the solutions to the relaxed problem are ones satisfying the equidistance constraint. These are then solutions to the fully constrained problem.

In the Appendix, we derive the following upper bounds on d^2:

$$d^2 \le \begin{cases} \frac{1}{2}q(q-1)^{-1}(g_{\max}-g_{\min})^2 T, & q \text{ even}, \quad (27a) \\ \frac{1}{2}(q+1)q^{-1}(g_{\max}-g_{\min})^2 T, & q \text{ odd}. \quad (27b) \end{cases}$$

An alternative and simpler derivation for q even is as follows. For any choice of g_s,

$$(g_i-g_j)^2 = (g_i-g_s)^2 - 2(g_i-g_s)(g_j-g_s) + (g_j-g_s)^2,$$

so that summing over i and j yields

$$\sum_{i=1}^{q}\sum_{j=1}^{q}(g_i-g_j)^2 = 2q\sum_{i=1}^{q}(g_i-g_s)^2 - 2q^2(c-g_s)^2,$$

where c is the centroid $c = q^{-1}\Sigma_{i=1}^{q} g_i$. Thus, from (18),

$$\begin{aligned} d^2 &= 2(q-1)^{-1}\int_0^T \sum_{i=1}^{q}\left[g_i(t)-g_s(t)\right]^2 dt \\ &\quad -2q(q-1)^{-1}\int_0^T\left[c(t)-g_s(t)\right]^2 dt \\ &\le 2(q-1)^{-1}\int_0^T \sum_{i=1}^{q}\left[g_i(t)-g_s(t)\right]^2 dt, \qquad (28) \end{aligned}$$

with equality holding if and only if $c(t) = g_s(t)$ for almost all $t \in [0, T]$. Taking $g_s(t) = \frac{1}{2}(g_{\max}+g_{\min})$ implies $|g_i(t) - g_s| \le \frac{1}{2}(g_{\max}-g_{\min})$, and the bound in (27a) then follows from (28). This bound holds for both odd and even values of q, but it is tight only for q even, and the more precise bound (27b) derived in the Appendix for q odd is the one that is achieved with equality.

Any set of q equidistant signals $\mathcal{G}$ satisfying (17) and achieving the upper bound (27a) for q even or (27b) for q odd is a signal set maximizing $R_{0,\infty}$. Signal sets having these properties can be identified for certain values of q by the following procedure. Partition $[0, T]$ into m equal subintervals, and define m functions $\rho_i(t)$, $i = 1, 2, \cdots, m$, that are piecewise constant having a constant value of one or zero over each subinterval. Then $\rho_i(t)$ can be identified by a binary codeword of length m bits. If we write $g_i^*(t) = g_{\min} + \rho_i(t)[g_{\max} - g_{\min}]$, it is enough to find q binary codewords of length m whose common Hamming distance satisfies the conditions in Table I. The last column in this table reflects a necessary condition for optimality that follows immediately from conditions for equality in (A1) that yields the upper bound (27); namely, for all $t \in [0, T]$,

TABLE I
CODE CONSTRAINTS

q	m	Hamming Distance	$\sum_{i=1}^{m} \rho_i(t)$
even	q-1	$\frac{1}{2}$q	$\frac{1}{2}$q
even	2(q-1)	q	$\frac{1}{2}$q
odd	q	$\frac{1}{2}$(q+1)	$\frac{1}{2}$(q-1)
odd	q	q+1	$\frac{1}{2}$(q-1)

i) for q even, $q/2$ of the signals take on value $g_{\max}$ and the remaining $q/2$ value $g_{\min}$,

ii) for q odd, $(q-1)/2$ or $(q+2)/2$ of the signals take on value $g_{\max}$ and the remainder $g_{\min}$.

This provides an additional check on the optimality of the following signal set and was an important aspect in our identification of it. For optimality, however, it is sufficient that the signal set be equidistant and achieve the appropriate upper bound (i.e., Hamming distance).

For q a multiple of four and such that a Hadamard matrix of order q exists, q codewords satisfying these conditions are easily obtained by deleting the first column (all ones) of the normalized Hadamard matrix [10], [11]. From this, $s = q - 1$ codewords satisfying the third row of Table I with s replacing q can be obtained by deleting the codeword of all ones. Also, $p = \frac{1}{2}q$ codewords satisfying the second row of Table I with p replacing q can be obtained by deleting all rows of the normalized Hadamard matrix that have a zero in (say) the second column and the deleting the first two columns. From this, $s = \frac{1}{2}q - 1$ codewords satisfying the fourth row of Table I with s replacing q can be obtained by deleting the codeword of all ones. Since Hadamard matrices for $q = 1, 2$, or a multiple of four are known up to $q = 200$ (and for many larger values), this procedure gives an optimizing signal set $\mathcal{G}$ for all $q \le 200$. Also several infinite families of Hadamard matrices are known, for example if $q = 2^k$ for some positive integer k: these coincide with cyclic maximal-length shift register codes, and they are also a subset of the first-order Reed–Muller codewords of this length.

We remark that complementation of an optimum signal set yields another optimum signal set. Also time-sharing of any two optimum signal sets yields another optimum signal set.

The average energy of these signal sets is easily calculated. If q is even, at any time $q/2$ signals have value $g_{\max}$ and the remainder $g_{\min}$, so

$$\bar{E}_g = \frac{1}{2}(g_{\max}^2 + g_{\min}^2)T, \qquad (29)$$

which implies

$$\bar{s} = \bar{E}_g - g_{\min}^2 T = \frac{1}{2}(g_{\max}^2 - g_{\min}^2)T. \qquad (30)$$

Similarly for q odd,

$$\bar{E}_g = \frac{1}{q}\left[\frac{1}{2}(q-1)g_{\max}^2 + \frac{1}{2}(q+1)g_{\min}^2\right]T, \tag{31}$$

which implies

$$\bar{s} = \bar{E}_g - g_{\min}^2 T = \frac{q-1}{2q}(g_{\max}^2 - g_{\min}^2)T. \tag{32}$$

This uses less energy than taking $(q+1)/2$ signals with value $g_{\max}$ and thus extends the range of average energies for which this choice is optimum; namely, the available average energy must exceed that required for $\bar{s}$ of (30) or (32), which yields condition (19c).

Finally, the distance d^* achieved by these signals that maximize $R_{0,\infty}$ when the peak-amplitude constraint predominates is given by

$$d^{*2} = \begin{cases} \dfrac{q}{q-1}\left[\dfrac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}, & q \text{ even,} \\[2ex] \dfrac{q+1}{q-1}\left[\dfrac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}, & q \text{ odd.} \end{cases} \tag{33}$$

Neither Constraint Predominates

If $\bar{s}_{\max}$ satisfies (19a) or (19c), the PPM signal set or, respectively, the Hadamard-derived signal set maximizes $R_{0,\infty}$. Unless $q=2$ or $q=3$, we are left with a range of values of $\bar{s}_{\max}$ for which a solution has yet to be identified. This "gap" region is specified in (19b). For $q=2$ or $q=3$, this region collapses to the empty set, and at the common upper limit of the range (19a) and lower limit of range (19c), the PPM or Hadamard-derived signal sets are equivalent and optimum. For $q \geq 4$, we now demonstrate that an optimum signal set results by time sharing the PPM and Hadamard-derived solutions.

The gap region has strictly positive length if $q \geq 4$, and then any point in either interval (19b) can be expressed as a strictly convex combination of the endpoints; that is, for q even and $\bar{s}_{\max}$ in the appropriate interval (19b), there exists a unique $\lambda \in (0,1)$ such that

$$\bar{s}_{\max} = \left[\frac{\lambda}{q} + \frac{(1-\lambda)}{2}\right](g_{\max}^2 - g_{\min}^2)T, \tag{34a}$$

while for q odd and $\bar{s}_{\max}$ in the appropriate interval (19b), there exists a unique $\lambda \in (0,1)$ such that

$$\bar{s}_{\max} = \left[\frac{\lambda}{q} + \frac{(1-\lambda)(q-1)}{2q}\right](g_{\max}^2 - g_{\min}^2)T. \tag{34b}$$

An optimum choice of modulation can now be given in terms of λ.

Lemma 3: For q even (respectively, odd) and $\bar{s}_{\max}$ in the appropriate interval specified in (19b), let λ be defined by (34a) (respectively, (34b)). Then an equidistant signal set that maximizes $R_{0,\infty}$ while satisfying the average-energy and peak-amplitude constraints with equality is for a fraction λ of the interval $[0,T]$ use the "full-width" PPM signal set (25) with $\epsilon = 1$ and T replaced by λT, and for a fraction $1-\lambda$ of $[0,T]$ use the signal set defined by the appropriate Hadamard matrix, as discussed in the previous section with T replaced by $(1-\lambda)T$.

Lemma 3 is proved as follows. For an arbitrary choice of $\alpha \in [0,1]$ and an arbitrary choice of maximum average energy $\bar{s}_{1,\max} \in [0,\bar{s}_{\max}]$ allocated to the interval $[0,\alpha T]$, we have from Lemma 1

$$\frac{1}{q(q-1)}\int_0^{\alpha T}\sum_{i=1}^{q}\sum_{j=1}^{q}\left[g_i(t)-g_j(t)^2\right]dt \leq 2\left[\frac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}_{1,\max}, \tag{35}$$

and from (27)

$$\frac{1}{q(q-1)}\int_{\alpha T}^{T}\sum_{i=1}^{q}\sum_{j=1}^{q}\left[g_i(t)-g_j(t)\right]^2 dt$$

$$\leq \begin{cases} \dfrac{(1-\alpha)Tq}{2(q-1)}(g_{\max}-g_{\min})^2, & q \text{ even,} \quad (36a) \\[2ex] \dfrac{(1-\alpha)T(q+1)}{2q}(g_{\max}-g_{\min})^2, & q \text{ odd.} \quad (36b) \end{cases}$$

Adding these expressions and using (18), we obtain

$$d^2 \leq \begin{cases} 2\left[\dfrac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}_{1,\max} + \dfrac{(1-\alpha)Tq}{2(q-1)}(g_{\max}-g_{\min})^2, \\ \quad q \text{ even,} \qquad (37a) \\ 2\left[\dfrac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}_{1,\max} \\ \quad + \dfrac{(1-\alpha)T(q+1)}{2q}(g_{\max}-g_{\min})^2, \\ \quad q \text{ odd.} \qquad (37b) \end{cases}$$

From Lemma 1, equality holds in (35) if and only if the following pair of conditions hold.

i) At any time $t \in [0,\alpha T]$, all signals take on the value $g_{\min}$, except at most one which takes on the value $g_{\max}$. This implies that the average energy $\bar{s}_1$ used on $[0,\alpha T]$ is

$$\bar{s}_1 = \frac{1}{q}\sum_{i=1}^{q}\int_0^{\alpha T} g_i^2(t)\,dt - g_{\min}^2 \alpha T \leq \frac{\alpha T}{q}(g_{\max}^2 - g_{\min}^2).$$

ii) $\bar{s}_1 = \bar{s}_{1,\max}$.

Thus a necessary condition for equality in (35) is

$$\bar{s}_{1,\max} \leq \frac{\alpha T}{q}(g_{\max}^2 - g_{\min}^2). \tag{38}$$

Furthermore, the derivation in the Appendix shows that equality holds in (36a) only if half of the signals take on the value $g_{\min}$ and the remainder $g_{\max}$. This involves an average energy usage $\bar{s}_2$ on $[\alpha T, T]$ of

$$\bar{s}_2 = \frac{1}{q}\sum_{i=1}^{q}\int_{\alpha T}^{T} g_i^2(t)\,dt - g_{\min}^2(1-\alpha)T = \frac{(1-\alpha)T}{2}(g_{\max}^2 - g_{\min}^2). \tag{39}$$

Because of the total average energy constraint $\bar{s}_1+\bar{s}_2<\bar{s}_{\max}$, we have, using (34a),

$$\bar{s}_1<\bar{s}_{\max}-\bar{s}_2=\left(\frac{\lambda}{q}+\frac{\alpha-\lambda}{2}\right)(g_{\max}^2-g_{\min}^2)T. \quad (40)$$

For equality to hold in (37a), it is necessary that both (35) and (36a) hold with equality, and necessary conditions for these are in turn (38) and, combining $\bar{s}_1=\bar{s}_{1,\max}$ with (40),

$$\bar{s}_{1,\max}<\left(\frac{\lambda}{q}+\frac{\alpha-\lambda}{2}\right)(g_{\max}^2-g_{\min}^2)T, \qquad q \text{ even}. \quad (41a)$$

For q odd, the corresponding necessary conditions for equality in (37b) become (38) and

$$\bar{s}_{1,\max}<\left[\frac{\lambda}{q}+\frac{(\alpha-\lambda)(q-1)}{2q}\right](g_{\max}^2-g_{\min}^2)T, \qquad q \text{ odd}. \quad (41b)$$

We now consider the selection of $\bar{s}_{1,\max}$ and α to maximize the upper bound (37a) subject to the constraints (38) and (41a), which are necessary conditions for it to hold with equality. Because both (38) and (41a) are constraints on $\bar{s}_{1,\max}$, we consider each in turn to be dominant in the sense of being more restrictive. The bound (38) is less than or equal to the bound (41a) if and only if $\alpha>\lambda$. Substituting (38) into (37a) and simplifying, we obtain

$$d^2<\frac{(g_{\max}-g_{\min})^2T}{2q(q-1)}\left[q^2-\alpha(q-2)^2\right].$$

Because we are considering $q>4$, this bound is maximized over $\alpha\in[\lambda,1]$ by the unique choice $\alpha=\lambda$. Similarly the bound (41a) is less than or equal to the bound (38) if and only if $\alpha<\lambda$. Substituting (41a) into (37a) and simplifying, we obtain

$$d^2<(g_{\max}-g_{\min})^2T\left[\frac{\alpha(q-2)}{2(q-1)}+\frac{q}{2(q-1)}+\frac{\lambda(2-q)}{q}\right].$$

Again because $q>4$, this bound is maximized over $\alpha\in[0,\lambda]$ by the unique choice $\alpha=\lambda$. Thus the bound (37a) is maximized, subject to the necessary conditions (38) and (41a), by taking $\alpha=\lambda$, and the corresponding maximum bound is

$$d^2<\left[\frac{2\lambda}{q}+\frac{q(1-\lambda)}{2(q-1)}\right](g_{\max}-g_{\min})^2T, \qquad q \text{ even}. \quad (42)$$

But this upper bound is readily achieved by the time-sharing of a "full width" PPM signal set for a fraction λ of $[0,T]$ and the Hadamard-derived signal set for the remaining fraction $1-\lambda$ of $[0,T]$. Furthermore the average energy required by this solution is exactly $\bar{s}_{\max}$. For q odd, a similar analysis leads again to the unique choice $\alpha=\lambda$ and the corresponding maximum bound

$$d^2<\left[\frac{2\lambda}{q}+\frac{(1-\lambda)(q+1)}{2q}\right](g_{\max}-g_{\min})^2T, \qquad q \text{ odd}. \quad (43)$$

This bound is achieved by the time-sharing of a full width PPM signal set for fraction λ of $[0,T]$ and the appropriate Hadamard-matrix-derived signal set for the remaining fraction $1-\lambda$ of $[0,T]$, and the average energy required by this solution is exactly $\bar{s}_{\max}$, as before.

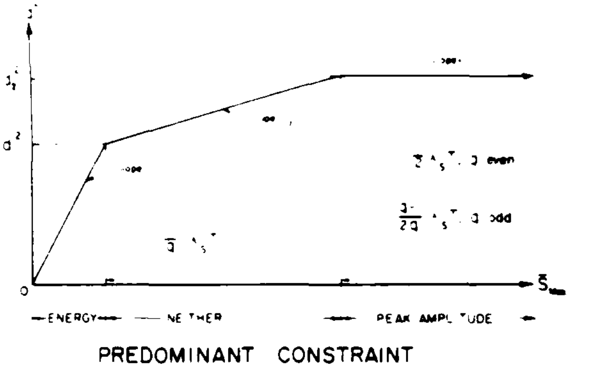


Fig. 2. d^2 for optimal signal sets.

IV. Efficient Energy Utilization

Denote by λ_s the count rate due to the signal alone when it is "on" for any of the optimal signal sets derived in the previous section. Then $g_{\max}^2=\lambda_s+\lambda_0$ and $g_{\min}^2=\lambda_0$, where λ_0 is the count rate due to the noise alone. In considering designs for efficient energy utilization we distinguish three situations depending on which of λ_s, $\bar{s}_{\max}$, and ϵ are adjustable and which are fixed. We seek to identify values of the adjustable parameters so that the cutoff rate per unit energy, $R_{0,\infty}/\bar{s}$, is greatest.

1) $\bar{s}_{\max}$ *adjustable*, λ_s *fixed:* The value of d^2 achieved with the optimal signal sets of the previous section is shown in Fig. 2 as a function of $\bar{s}_{\max}$, assuming that λ_s is a fixed constant. Here d^2 is a piecewise linear function of $\bar{s}_{\max}$ with the following parameters:

$$d_1^2=\frac{2}{q}\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2\lambda_0T, \quad (44)$$

$$d_2^2=\begin{cases}\dfrac{q}{2(q-1)}\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2\lambda_0T, & q \text{ even},\\[2ex] \dfrac{q+1}{2(q-1)}\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2\lambda_0T, & q \text{ odd},\end{cases} \quad (45)$$

$$s_1=2(\lambda_0/\lambda_s)\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2. \quad (46)$$

$$s_2=\begin{cases}\dfrac{q-2}{q-1}(\lambda_0/\lambda_s)\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2, & q \text{ even},\\[2ex] \dfrac{q^2-3q+4}{q^2-4q+3}(\lambda_0/\lambda_s)\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2, & q \text{ odd}.\end{cases} \quad (47)$$

Using (10), with equality for optimal signal sets, and using the expressions for d^2 implied by Fig. 2 and (44)–(46), we conclude that

$$dR_{0,\infty}/d\bar{s}_{\max}=0.72(q-1)\left[(q-1)+\exp\left(\frac{1}{2}d^2\right)\right]^{-1} \text{ (slope)}, \quad (48)$$

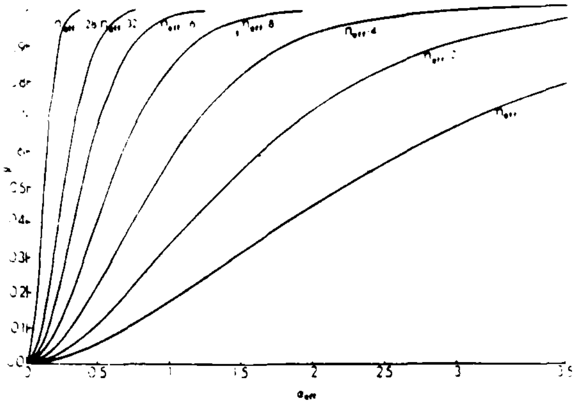


Fig. 3. Cutoff rate as a function of signal-to-noise ratio.

where the factor "slope" is s_1, s_2, or zero depending on which constraint, if any, predominates. Thus $dR_{0,\infty}/d\bar{s}_{\max}$ decreases monotonically with increasing $\bar{s}_{\max}$, so the signal energy is used most efficiently when $\bar{s}_{\max}$ is small, where the energy constraint predominates and where the PPM signal set is optimal and utilizes energy $\bar{s}=\bar{s}_{\max}$. This situation is analogous to that studied by Massey [8], [9] for the additive Gaussian noise channel. For $\bar{s}=\bar{s}_{\max}$ small, we conclude that

$$R_{0,\infty}/\bar{s} \approx 1.44\frac{q-1}{q}(\lambda_0/\lambda_s)\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2, \quad (49)$$

with equality for $\bar{s}=0$. Hence

$$R_{0,\infty} < 1.44\frac{q-1}{q}(\lambda_0/\lambda_s)\left[1-(1+\lambda_s/\lambda_0)^{1/2}\right]^2\bar{s} \quad (50)$$

is an upper bound on $R_{0,\infty}$ for any choice of $\bar{s}$ and any choice of modulation with near equality holding when $\bar{s}$ is small and for the PPM signal set. Since $\bar{s}=\epsilon T\lambda_s/q$ for the PPM signal set, this means that when λ_s is fixed, the most efficient energy usage occurs for narrow pulses, ϵ being selected to be as small as is practically feasible.

2) *λ_s adjustable, $\bar{s}$ fixed:* By a somewhat messy but straightforward calculation it is readily verified that $dR_{0,\infty}/d\lambda_s > 0$ for $\bar{s}$ fixed. Thus $R_{0,\infty}$, and hence $R_{0,\infty}/\bar{s}$ for $\bar{s}$ fixed is a nondecreasing function of λ_s. Consequently the most efficient energy usage is achieved by selecting λ_s large and operating in the region where the energy constraint predominates. This implies using the PPM signal set with as large a value of signal count rate λ_s as practical and sufficiently narrow pulses that $\bar{s}=\epsilon T\lambda_s/q$.

3) *$\bar{s}$ adjustable, PPM signal set with ϵ fixed:* The PPM signal set with a fixed pulsewidth $\epsilon T/q$ maximizes $R_{0,\infty}$ provided the energy constraint predominates, which we assume. For ϵ fixed and $\bar{s}=\epsilon T\lambda_s/q$, we find that

$$R_{0,\infty}(\bar{s},\bar{n}_{eff}) = \log_2 q - \log_2\Bigg\{1+(q-1) \cdot \exp\left[-\frac{1}{q}\bar{n}_{eff}\left(1-\sqrt{1+q\alpha_{eff}}\right)^2\right]\Bigg\}, \quad (51)$$

where we define

$$\bar{n}_{eff} = \epsilon\bar{n} = \epsilon\lambda_n T$$

to be the "effective" average number of noise counts per channel use and where

$$\alpha_{eff} = \bar{s}/\bar{n}_{eff} \quad (52)$$

is the signal-to-noise energy ratio. Graphs of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ as a function of α_{eff} for several values of $\bar{n}_{eff}$ are given in Fig. 3. These graphs are seen to increase monotonically with α_{eff} for each fixed value of $\bar{n}_{eff}$. Thus as expected the performance improves systematically for fixed $\bar{n}_{eff}$ as the average signal energy per channel use $\bar{s}$ increases. However, although starting from $\bar{s}=0$ the performance initially improves rapidly, there is a point of diminishing returns after which there is only marginal improvement for further increases in $\bar{s}$. For each $\bar{n}_{eff}$, there is an $\bar{s}=\bar{s}^*(\bar{n}_{eff})$ such that for all $\bar{s}>0$ there holds

$$\frac{R_{0,\infty}(\bar{s},\bar{n}_{eff})}{\bar{s}} < \frac{R_{0,\infty}(\bar{s}^*,\bar{n}_{eff})}{\bar{s}^*}. \quad (53)$$

This value of $\bar{s}$ can be found graphically for each $\bar{n}_{eff}$ by pivotting a vertical line about the origin ($R_{0,\infty}=0$, $\alpha_{eff}=0$) in Fig. 3 until it lies tangent to the graph of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$. The abscissa of the point of tangency is $\bar{s}^*/\bar{n}_{eff}$. Inequality (53) holds because the graph of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ lies on or below the line so constructed for all $\alpha_{eff}>0$. It follows from (53) that the most efficient usage of energy, in the sense that the cutoff rate per unit energy is greatest, is achieved when $\bar{s}=\bar{s}^*$. The dashed line in Fig. 3 is a fit of $\bar{s}^*/\bar{n}_{eff}$ versus $\bar{n}_{eff}$ obtained graphically by connecting together the points of tangency described above. From this fit we find for the range of average noise counts in the figure that $\bar{s}^*$ and $\bar{n}_{eff}$ are approximately related by the following power law:

$$\bar{s}^* \approx 2.349\bar{n}_{eff}^{0.452}. \quad (54)$$

This is shown as the solid line in Fig. 4. A measure of the range of energies nearly as efficient as $\bar{s}^*$ can be determined for each $\bar{n}_{eff}$ from the values of $\alpha_{eff}=\bar{s}/\bar{n}_{eff}$ in Fig. 3 for which $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ is close to, say within ten percent of, the ordinate of the line of tangency constructed as above. Values of $\bar{s}$ within the dashed lines in Fig. 4 satisfy this ten percent condition; Fig. 4 implies that for maximally efficient energy utilization $\bar{s}$ should be kept within about ±2 dB of $\bar{s}^*$.

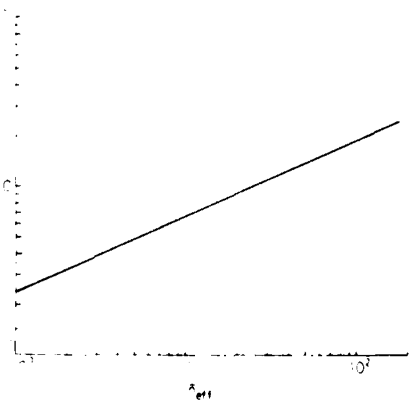


Fig. 4. Optimal signal energy as a function of noise energy per channel use

V. Effect of Finite Output Quantization

The cutoff rate decreases from $R_{0,\infty}$ as the dimension q' of the output alphabet decreases. This degradation is greatest for a binary input alphabet ($q=2$) when $q'=2$, which corresponds to making bit by bit decisions without any coding. For a Gaussian model Massey [8], [9] concludes that choosing $q'=2$ results in a quantization loss of 2.04 dB; that is, in the efficient range of energy utilization for the Gaussian model, the energy per channel use must be about 2 dB greater for $q'=2$ in order to achieve the same cutoff rate as when $q'=\infty$. Massey also concludes that for $q'=8$ there is virtually no quantization loss. The degradation for the Poisson model is somewhat smaller than that found by Massey when $5<\bar{n}_{\text{eff}}<40$ and is of about the same order when $\bar{n}_{\text{eff}}=1$.

Suppose the input and output alphabets are $\mathcal{X}=\{0,1\}$ and $\mathcal{Y}=\{0,1\}$, so that $q=q'=2$. We adopt a binary pulse-position modulation format with pulses of duration $\epsilon T/2$ because this maximizes $R_{0,\infty}$ when the energy constraint predominates. For this choice each symbol interval is divided into two equal subintervals and output letters are generated according to "produce one if $n[0,\epsilon T/2]<n[T/2,(1+\epsilon)T/2]$, otherwise produce zero," where $n[0,\epsilon T/2]$ and $n[T/2,(1+\epsilon)T/2]$ are the number of points observed in the first and second signalling interval, respectively. Here $n[0,\epsilon T/2]$ and $n[T/2,(1+\epsilon)T/2]$ are independent Poisson random variables with mean parameters $\bar{s}+(\bar{n}_{\text{eff}}/2)$ and $\bar{n}_{\text{eff}}/2$, respectively, when zero is the input letter and $\bar{n}_{\text{eff}}/2$ and $\bar{s}+(\bar{n}_{\text{eff}}/2)$, respectively, when one is the input letter. As in the previous discussion $\bar{s}$ and $\bar{n}_{\text{eff}}$ are the average number of signal counts received per channel use and effective number of noise counts received per channel use. It is straightforward to conclude for these assumptions that the cutoff rate is given by

$$R_{0,q'=2}=1-\log_2\{1+2[p(1-p)]^{1/2}\}. \tag{55}$$

where p is the binary error probability associated with producing an output symbol 1 (or a 0) when the input symbol is a zero (or a one, respectively). This error probability is given graphically for certain values of $\bar{n}_{\text{eff}}$ and a range of $\bar{s}$ by Pratt [4, p. 209: identify $\bar{n}_{\text{eff}}=2\mu_{H,B}$ and $\bar{s}=\mu_{S,B}$].

The values tabulated in Table II were obtained as follows: 1) $\bar{s}^*$ is obtained from (54) for each $\bar{n}_{\text{eff}}$; 2) p^* is the value of p in (55) such that $0<p^*<1$ and $R_{0,q'=2}=R^*_{0,\infty}$; and 3) $\bar{s}$ is obtained by interpolation from the graph given by Pratt. Thus $\bar{s}^*$ and $\bar{s}$ of the table yield the same cutoff rate for $q'=\infty$ and $q'=2$, respectively. To within the accuracy that the interpolation step can be accomplished, we conclude that about 1.5 dB more signal energy is required with hard decisions than with infinitely soft decisions for $\bar{n}_{\text{eff}}$ in the range of five to 40 counts per channel use.

TABLE II
Degradation due to Finite Quantization

$\bar{n}_{\text{eff}}$	$\bar{s}^*$	$R^*_{0,\infty}$	p^*	$\bar{s}$	$10\log(\bar{s}/\bar{s}^*)$
1	2.35	0.53	0.038	3.8	2.09
5	4.86	0.65	0.020	7.0	1.58
10	6.65	0.68	0.016	9.25	1.43
20	9.10	0.70	0.014	12.7	1.45
40	12.45	0.71	0.013	16.9	1.33

VI. Effect of Input Alphabet Dimension

For an input alphabet of dimension $q, q'=\infty$, and an average energy constraint that predominates, $q$ary pulse-position modulation maximizes the cutoff rate. We now consider the effect of q in each of the three situations identified in Section IV.

1) *$\bar{s}_{\max}$ adjustable, λ_s fixed:* From (48) and (49), increasing q from two to infinity implies that the greatest rate per unit energy that can be achieved increases by a factor of two. Moreover, examination of the graphs of $R_{0,\infty}/\bar{s}$ for $R_{0,\infty}$ given by (10) with equality and with $d^2=s_1\bar{s}$, where s_1 is given in (46), shows that the range of values of $\bar{s}$ for which the approximation (49) holds closely increases as q increases; in other words, the range of efficient signal energies is extended as q is increased.

2) *λ_s adjustable, $\bar{s}$ fixed:* For λ_s large and the PPM signal set, we see from Fig. 2 and (46) that $d^2 \simeq 2\bar{s}$. Then

$$R_{0,\infty} \approx \log_2 q - \log_2[1+(q-1)e^{-\bar{s}}] < \frac{q-1}{q}\bar{s}. \tag{56}$$

Hence, for large λ_s, $R_{0,\infty}/\bar{s} \lesssim (q-1)/q$ and therefore the largest rate per unit energy increases by no more than a factor of two as q increases from two to infinity.

3) *$\bar{s}$ adjustable, PPM signal set with ϵ fixed:* A graph of (51) as a function α_{eff} for $\bar{n}_{\text{eff}}=16$ and various values of q is shown in Fig. 5. For each q, there is a corresponding

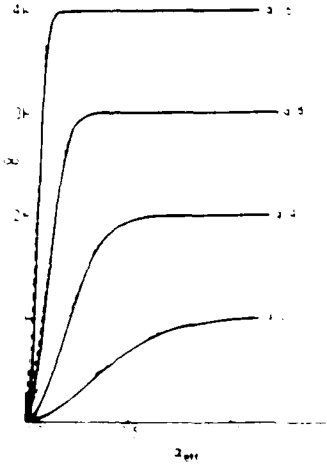


Fig. 5. Effect of input-alphabet dimension on cutoff rate.

signal energy that is most efficient; this can be found graphically in the same manner as before, as indicated by the lines of tangency. These efficient energies depend upon q; very roughly we find from the graphs that $(\bar{s}^*/16)q \approx 1$, so that the most efficient signal energy decreases as q increases. This implies a significant potential improvement in performance at low signal energies by using a large input-alphabet dimension q and $q$ary pulse-position modulation. These observations appear to hold for other values of $\bar{n}_{\text{eff}}$ as well.

VII. Effect of Random Detector Gain

Let $\{M(t);\ t>0\}$ be a compound Poisson counting process defined by

$$M(t)=\sum_{n=0}^{N(t)} u_n, \tag{57}$$

where $\{N(t);\ t>0\}$ is the Poisson counting process defined above, $u_0=0$, and $\{u_n;\ n=1,2,\cdots\}$ is a sequence of independent identically distributed random variables each having an integer value greater than zero. Here $\{N(t);\ t>0\}$ models primary photoelectron conversions, and u_n models the number of secondary electrons appearing at the detector output due to the nth conversion. This random gain is an important effect encountered, for example, with avalanche detectors used in optical-fiber communication systems.

In considering a digital-data communication system in which measurements are derived from $\{M(t);\ t>0\}$, it is of interest to known the cutoff rate $R_{0,\infty}$ for infinitely fine quantization. As before, this quantity then places an upper limit on the performance of any receiver employing finite quantization, such as an "integrate dump" receiver [12], [13] in which $M(nT)-M[(n-1)T]$ is used to make a decision about the nth transmitted symbol.

We find $R_{0,\infty}$ to be identical to that in (6), so random detector-gain neither degrades nor enhances the cutoff rate for infinitely fine output quantization. This is because the distribution of the random gains is unaffected by the choice of transmitted signal on our model and can be verified mathematically by the following steps. First we write the summation over k in (3) as $f(i,j)=E_j[\Lambda_{i,j}^{1/2}(y)]$, where

$$\Lambda_{i,j}(Y)=p_{y|x}(Y|X_i)/p_{y|x}(Y|X_j) \tag{58}$$

is the likelihood ratio for symbol X_i relative to symbol X_j and $E_j(\cdot)$ denotes a conditional expectation given X_j. As the output quantization is refined, this becomes

$$f(i,j)=E_j\left[\Lambda_{i,j}^{1/2}(M(t);\ 0<t<T)\right], \tag{59}$$

where $\Lambda_{i,j}(M(t);\ 0<t<T)$ is given by the ratio of the sample function densities [18] of $\{M(t);\ 0<t<T\}$ for symbols X_i and X_j. This likelihood ratio is found not to be a function of the random gains, and the assertion follows.

A consequence of this assertion is that many of the conclusions reached in the preceding sections also apply in the presence of random detector-gain. At the present time there are too few published results on the binary error-probability for an integrate-and-dump receiver for us to examine the potential benefits of employing finer output quantization, but this is a matter of some practical interest for fiber-optic systems.

VIII. Polarization Modulation

Suppose that binary orthogonal polarization modulation can also be employed in the optical modulator of Fig. 1 in addition to temporal modulation. Then the scalar field $E(t,r)$ becomes a vector $(E_1(t,\bar{r}),E_2(t,\bar{r}))$ in which one component is the 0° field and the other one the 90° field. A polarization decomposition of the received field followed by direct detection in each channel then results in two independent point processes, which we label $N_1(t)$ and $N_2(t)$, $0<t<T$. Assume that when the input code letter is $X_i \in \{X_1,X_2,\cdots,X_q\}$ that the count rate for $N_1(t)$ is

$$\lambda_{1i}(t)=s_{1i}(t)+\lambda_0=g_{1i}^2(t), \tag{60a}$$

and for $N_2(t)$ is

$$\lambda_{2i}(t)=s_{2i}(t)+\lambda_0=g_{2i}^2(t). \tag{60b}$$

Following the procedure used in the last section, as $q'\to\infty$, the sum over k in (3), call it $f(i,j)$, becomes

$$\begin{aligned} f(i,j)&=E_j\left\{\Lambda_{i,j}^{1/2}[N_1(t),N_2(t);\ 0<t<T]\right\}\\ &=\exp\left(-\frac{1}{2}d_{ij}^2\right),\end{aligned} \tag{61}$$

where

$$d_{ij}^2=\int_0^T\left([g_{1i}(t)-g_{1j}(t)]^2+[g_{2i}(t)-g_{2j}(t)]^2\right)dt. \tag{62}$$

The steps leading to (10) remain unchanged with (62) replacing (8).

We now assume that each of the signals in $\mathcal{S}_1=\{E_{11}(t,\bar{r}),E_{12}(t,\bar{r}),\cdots,E_{1q}(t,\bar{r})\}$ and $\mathcal{S}_2=\{E_{21}(t,\bar{r}),E_{22}(t,\bar{r}),\cdots,E_{2q}(t,\bar{r})\}$ satisfy the average-energy and peak-amplitude constraints in the section about modulator design based on $R_{0,\infty}$. Then we have the following optimization problem: select signals in $\mathcal{G}_1=\{g_{11}(t),g_{12}(t),\cdots,g_{1q}(t)\}$ and $\mathcal{G}_2=\{g_{21}(t),g_{22}(t),\cdots,q_{2q}(t)\}$ to maximize

$$\begin{aligned} d^2=[q(q-1)]^{-1}\sum_{i=1}^{q}\sum_{j=1}^{q}\int_0^T&\left([g_{1i}(t)-g_{1j}(t)]^2\right.\\ &\left.+[g_{2i}(t)-g_{2j}(t)]^2\right)dt\end{aligned} \tag{63}$$

subject to the following constraints.

i) *Equidistance constraint:* The quantities in (62) should be the same whenever $i\neq j$.

ii) *Average-energy constraint:* (15) should be satisfied for both signal sets $\mathcal{G}_1$ and $\mathcal{G}_2$.

iii) *Peak-amplitude constraint:* (17) should be satisfied for both signal sets $\mathcal{G}_1$ and $\mathcal{G}_2$.

By paralleling the development leading to Lemma 1, we have the following.

Lemma 1′: Let $\bar{s}_{\max}$ be the maximum average signal counts per channel use in each polarization component, and suppose (17) applies to both signal sets $\mathcal{G}_1$ and $\mathcal{G}_2$. Then

$$d^2 < 4\left[\frac{g_{\max}-g_{\min}}{g_{\max}+g_{\min}}\right]\bar{s}_{\max}. \tag{64}$$

Furthermore, equality holds if and only if both a) at any time $t \in [0,T]$, all signals in $\mathcal{G}_1$ and $\mathcal{G}_2$ take on the value $g_{\min}$ except at most one in $\mathcal{G}_1$ and one in $\mathcal{G}_2$ that takes on the value $g_{\max}$; and b)

$$\frac{1}{q}\sum_{i=1}^{q}\int_0^T g_{ki}^2(t)\,dt - g_{\min}^2 T = \bar{s}_{\max}, \qquad \text{for } k=1,2.$$

A signal set $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ that is equidistant in the sense that the quantities in (62) are the same whenever $i \neq j$, achieves the upper bound in Lemma 1′ with equality, and which therefore maximizes $R_{0,\infty}$ when the average-energy constraint predominates in each polarization component, is characterized in the following lemma for q even.

Lemma 2′: If q is even and $\bar{s}_{\max}$ satisfies (19a), equality is achieved in (64) by the following signal set. For $1 < i < (q/2)$ and $j = i + (q/2)$,

$$g_{1i}^*(t) = g_{2j}^*(t)$$
$$= \begin{cases} g_{\max}, & (i-1)2T/q < t < (i-1+\epsilon)2T/q, \\ g_{\min}, & \text{otherwise for } 0 < t < T, \end{cases}$$
$$g_{1j}^*(t) = g_{2i}^*(t) = g_{\min}, \qquad 0 < t < T,$$

where ϵ is given in (26).

The verification of Lemma 2′ parallels that of Lemma 2. It is interesting to note that for $q = 4$ this signal set, then called "quaternary pulse modulation," is used in the one gigabit per second optical communication system reported by M. Ross *et al.* [15].

Acknowledgment

We particularly thank Professor J. L. Massey of the University of California, Los Angeles, for several enlightening discussions about the cutoff rate criterion. Professors S. Mitter (Massachusetts Institute of Technology) and M. Pursley (University of Illinois) and Dr. M. Hurtado (IBM Corporation) also provided useful insights into various parts of the development. We are indebted to Professor R. Kennedy of the Massachusetts Institute of Technology for identifying an error in an earlier version of the manuscript.

Appendix
Derivation of (27)

Let $\mathcal{Q} = \{1, 2, \cdots, q\}$ and define J^* by

$$J^* = \max_{g_k(\cdot),\, k \in \mathcal{Q}} \int_0^T \sum_{i,j \in \mathcal{Q}} \left[g_i(t) - g_j(t)\right]^2 dt. \tag{A1}$$

Then

$$J^* < T \max_{t \in [0,T]} \max_{g_k(t),\, k \in \mathcal{Q}} \sum_{i,j \in \mathcal{Q}} \left[g_i(t) - g_j(t)\right]^2, \tag{A2}$$

with equality if and only if the integrand in (A1) is a constant independent of t. Thus we consider the problem of choosing q real numbers g_k, $k \in \mathcal{Q}$ to maximize

$$I(g) = \sum_{i,j \in \mathcal{Q}} (g_i - g_j)^2 \tag{A3}$$

subject to $g_{\min} < g_i < g_{\max}$, $i \in \mathcal{Q}$. A necessary condition for g_i^*, $i \in \mathcal{Q}$ to minimize $-I$ (and so maximize I) is the existence of $2q$ real numbers $\nu_i > 0$, $\mu_i > 0$, $i = 1, 2, \cdots, q$, such that [16][1]:

$$L(g^*, \mu, \nu) < L(g, \mu, \nu), \qquad \text{for all } g \text{ in } R^q, \tag{A4}$$

$$g_i^* < g_{\max} \text{ implies } \mu_i = 0, \tag{A5}$$

$$g_i^* > g_{\min} \text{ implies } \nu_i = 0, \tag{A6}$$

where the Lagrangian L is defined by

$$L(g,\mu,\nu) = -\sum_{i,j \in \mathcal{Q}} (g_i - g_j)^2 + \sum_{i \in \mathcal{Q}} \mu_i (g_i - g_{\max}) + \sum_{i \in \mathcal{Q}} \nu_i (g_{\min} - g_i). \tag{A7}$$

Setting the derivative of L with respect to g_i equal to zero, we obtain for $i \in \mathcal{Q}$

$$-2qg_i^* + 2qc^* + \mu_i - \nu_i = 0, \tag{A8}$$

where $c^* = q^{-1}\Sigma_{i \in \mathcal{Q}} g_i$. From (A5) and (A6), if $g_{\min} < g_i^* < g_{\max}$, then $\mu_i = \nu_i = 0$, and

$$g_i^* = c^*. \tag{A9}$$

Thus each g_i^* takes on one of three values: $g_{\min}$, $g_{\max}$, or c^*. Let there be $n_{\min}$, $n_{\max}$, and $q - n_{\min} - n_{\max}$ of these, respectively. From the definition of c^* we have

$$qc^* = n_{\min} g_{\min} + n_{\max} g_{\max} + (q - n_{\min} - n_{\max})c^*$$

or

$$c^* = (n_{\min} g_{\min} + n_{\max} g_{\max})/(n_{\min} + n_{\max}). \tag{A10}$$

Furthermore,

$$\sum_{i,j \in \mathcal{Q}} (g_i^* - g_j^*)^2 = 2n_{\min} n_{\max} (g_{\max} - g_{\min})^2 + 2n_{\max}(q - n_{\min} - n_{\max})(g_{\max} - c^*)^2 + 2n_{\min}(q - n_{\min} - n_{\max})(c^* - g_{\min})^2. \tag{A11}$$

Substituting (A10) into (A11) and simplifying, we obtain

$$I(g^*) = L(g^*, \mu, \nu) = 2q(g_{\max} - g_{\min})^2 \left(\frac{1}{n_{\min}} + \frac{1}{n_{\max}}\right)^{-1}, \tag{A12}$$

and this is to be minimized subject to $0 < n_{\min} + n_{\max} < q$, $n_{\min}$ and $n_{\max}$ being nonnegative integers, which is the same as the minimization of $(1/n_{\min}) + (1/n_{\max})$ with the same constraints. The solution to this is for q even, $n_{\min} = n_{\max} = q/2$; and for q

[1] This is a necessary condition for a *regular* point g^* to minimize $-I$ subject to $g_{\min} < g_i < g_{\max}$, $i \in \mathcal{Q}$. Since $g_i - g_{\max} < 0$ and $g_{\min} - g_i < 0$ cannot be simultaneously active (that is, satisfied with equality), it is evident that the set of gradient vectors of the active constraints can include e_i (the ith natural basis vector) or $-e_i$ but not both (and possibly neither). *Thus the set of gradient vectors of the active constraints is* linearly independent for any g, and any g is therefore regular.

odd, either $n_{\min}=(q+1)/2$ and $n_{\max}=(q-1)/2$ or $n_{\min}=(q-1)/2$ and $n_{\max}=(q+1)/2$. Thus[2] for q even we have

$$I(g^*)=\frac{1}{2}q^2(g_{\max}-g_{\min})^2 \tag{A13}$$

and for q odd,

$$I(g^*)=\frac{1}{2}(q+1)(q-1)(g_{\max}-g_{\min})^2. \tag{A14}$$

The corresponding upper bounds on $d^{*2}=q^{-1}(q-1)J^*$ are, from (A1),

$$d^{*2}<\begin{cases}\frac{1}{2}q(q-1)(g_{\max}-g_{\min})^2T, & q \text{ even,}\\ \frac{1}{2}(q+1)q^{-1}(g_{\max}-g_{\min})^2T, & q \text{ odd.}\end{cases} \tag{A15}$$

[2] Because this is the only solution to the necessary conditions (A4)–(A6), either it is the maximum of I or none exists. But, by the Weierstrass theorem, the continuous function I defined by (A3) achieves its maximum on the compact subset of R^q defined by $g_{\min} < g_i < g_{\max}$. Thus it is indeed the maximum.

References

[1] S. Karp, E. L. O'Neill, and R. M. Gagliardi, "Communication theory for the free space optical channel," *Proc. IEEE*, vol. 58, pp. 1611–1626, Oct. 1970.

[2] E. V. Hoversten, R. O. Harger, and S. J. Halme, "Communication theory for the turbulent atmosphere," *Proc. IEEE*, vol. 58, pp. 1626–1650, Oct. 1970.

[3] E. V. Hoversten, "Optical communication theory," in *Laser Handbook*, F. T. Arecchi and E. O. Schulz-DuBois, Eds. Amsterdam: North-Holland, 1972, ch. F8.

[4] W. K. Pratt, *Laser Communication Systems*. New York: Wiley, 1969.

[5] S. Karp and R. M. Gagliardi, *Optical Communications*. New York: Wiley, 1976.

[6] J. M. Wozencraft and R. S. Kennedy, "Modulation and demodulation for probabilistic coding," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 291–297, July 1966.

[7] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[8] J. L. Massey, "Coding and modulation in digital communications," in *Proc. Int. Zurich Sem. on Digital Communications*, Zurich, Switzerland, Mar. 12–15, 1974.

[9] ——, Course notes for EE 453, Department of Elec. Eng., Univ. of Notre Dame, Notre Dame, IN, 1976.

[10] S. W. Golomb, Ed., *Digital Communications*. Englewood Cliffs, NJ: Prentice-Hall, 1964, p. 53.

[11] W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*. Cambridge, MA: M.I.T., 1972, sec. 5.6.

[12] S. D. Personick, P. Balaban, J. H. Bobsin, and P. R. Kumar, "A detailed comparison of four approaches to the calculation of the sensitivity of optical fiber system receivers," *IEEE Trans. Commun.*, vol. COM-25, pp. 541–548; May 1977.

[13] J. E. Mazo and J. Salz, "On optical data communication via direct detection of light pulses," *Bell Syst. Tech. J.*, vol. 55, pp. 347–369, Mar. 1976.

[14] D. L. Snyder, *Random Point Processes*. New York: Wiley, 1975, p. 156.

[15] M. Ross, P. Freedman, J. Abernathy, G. Matassov, J. Wolf, and J. D. Barry, "Space optical communications with the Nd:YAG Laser," *Proc. IEEE*, vol. 66, pp. 319–344, Mar. 1978.

[16] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. New York: Wiley, 1969.

Reprint of Paper:

"Some Informationally-Decentralized Network Algorithms", Jeffrey M. Abram and Ian B. Rhodes, Proceedings of the 1980 Joint Automatic Control Conference, San Francisco, California, August 13-15, 1980

SOME INFORMATIONALLY-DECENTRALIZED NETWORK ALGORITHMS

Jeffrey M. Abram

Ian B. Rhodes
Department of Electrical and Computer Engineering
University of California. Santa Barbara
Santa Barbara. California

ABSTRACT

Several decentralized algorithms for determining the shortest paths in a network are presented. Both static and dynamic networks are discussed. Each algorithm has localized information and communication requirements, operates asynchronously, and converges to the optimal solution in finite time.

INTRODUCTION

This paper concerns shortest path algorithms that enable each node in a network to calculate its shortest distance to any other node using only local knowledge of the network topology and only local information transfer between adjacent nodes. Such algorithms are of obvious importance in many applications.

Our initial algorithm [1] was developed for a static network in which branch lengths and topology remain constant, though it can accommodate limited changes. We discuss here a number of modifications to the algorithm that enable it to operate in a dynamic network in which branch lengths can increase or decrease, and nodes or branches can be added to or removed from the network. The ability of an algorithm to handle such topological changes is essential in most practical applications.

THE STATIC ALGORITHM

We present a brief outline of the algorithm described in [1]. Consider a directed graph with branch lengths unrestricted in sign, but the sum of the lengths in any closed loop of the network is assumed to be positive. For convenience, we assume that the existence of a link from one node to another implies the existence of the opposite link. Very little topological information is needed. Each node needs to know only which nodes are its neighbors, the length of the branch to each of its neighbors, and distance information received from these neighbors. For each ultimate destination, a node calculates and stores an assessment of the shortest distance via each of its neighbors, the smallest of these is taken to be its assessment of the distance to that destination. Also stored is the identity of the neighbor that achieves this minimal distance.

Whenever a node's current shortest distance to a destination changes, either through reinitialization or as a result of new information received from a neighbor, this new distance is transmitted to all neighbors. At the conclusion of the algorithm, each node will know the shortest distance to each other node (or that no path exists), the next node in the path that achieves this distance, and the shortest distance via each alternative neighbor. The algorithm is guaranteed to converge, even if it is implemented in an asynchronous manner.

DYNAMIC NETWORK ALGORITHMS

While the above algorithm handles static networks, many problems arise for a dynamic network model. These phenomena include branch lengths decreasing and increasing, branches being introduced into the network, and branches failing or being removed from the network. The static algorithm can guarantee convergence for branch length decreases, but not increases or branch failures.

For this reason, we have developed several modified versions of the static algorithm, two of which are described in this paper. Each of these modified algorithms is based primarily on some form of reinitialization of the basic static algorithm; they differ mainly in the mechanics of the reinitialization. Both share the convergence properties of the static algorithm, provided that the topology of the network remains constant long enough for the algorithms to find the new solution.

Local Reinitialization

It is not sufficient to merely disseminate a reinitialization command throughout the network when a branch length increase or branch failure occurs. Once a node has reinitialized, it needs some guarantee that subsequently-received information is based on all affected nodes having reinitialized also. Thus, a mechanism has been devised for each node to determine which distance information it receives is trustworthy (in that all nodes further down the corresponding path are guaranteed to have reinitialized) and which is not. In simple terms, on hearing that a hranch length increase or failure has taken place, a node ignores distance information sent by any questionable neighbor until that neighbor acknowledges that it, too, is aware of the change. As each neighbor in turn so acknowledges, the embargo on its information is removed. In this way, some convergence toward the new solution can be taking place while news of the change is still propagating through the network.

More precisely, this algorithm involves the use of a "special action", which is initiated by a node detecting a branch increase or failure. When this occurs, the initiator assigns a unique index to the special action (consisting of his node number and a counter), and does the following:

1. Reinitializes
2. Places an embargo, indexed by the special action, on distances received from every neighbor.
3. Transmits all shortest distances to each neighbor, along with the special action index.

Each neighbor receiving this information takes analogous action, in Step 2 placing an embargo on all his neighbors except for the one he has just heard from, and in Step 3 using the same special action index. He then waits until a message is received from a neighbor, at which point his action is governed by the following:

Case 1. Message contains no special action index.
- A. If there is no embargo on this node, calculate the new distances via this neighbor and proceed normally.
- B. Else, ignore the message.

Case 2. Message contains a special action index.
- A. If there is no embargo on this node with the same index as in the message, perform steps 1 - 3 above.
- B. Else, remove the matching ban from this particular neighbor. If no other embargo exists, calculate new distance via this node. Otherwise, ignore distance component of message.

This algorithm can effectively handle increasing and decreasing branch lengths and failures, but some other technique must be introduced if one wishes to add nodes or links to the network. One such technique is described here.

Global Reinitialization Procedure. This procedure brings about global reinitialization by effectively suspending all communication of distance information for a sufficiently long period of time to insure that all nodes have reinitialized. Several possible mechanisms for achieving this present themselves: one is for one of the nodes incident to the new link to decide upon a future time at which communication of distance information based on reinitialization will resume, and to send this to his neighbors who continue to propagate it throughout the network. Implicit here is the existence of a time base common to all nodes, and the availability to each node of (at least an upper bound on) the time it takes for the "reinitialization message" he initiates to propagate throughout the network, which implies a more global knowledge of the network.

The resulting algorithm uses primarily localized reinitialization, with global reinitialization being used, hopefully infrequently, only to allow new links (nodes) to be added to the network. The algorithm can accommodate an arbitrary number of topological changes, requiring only an eventual cessation of topological changes in order to converge to the optimal solution.

Local Reinitialization with Acknowledgments

This version of the algorithm is a modification of the previous one, developed to solve the problem of adding 'inks. If a node remembers the indices of special actions after it processes them, it can insure that any new neighbors that it acquires are informed of the most recent special actions. But how long must a node remember which actions have occurred? In a network in which branch increases are a common occurrence, it would be infeasible for a node to remember every special action that it had received. Actually, a node need remember a particular special action only until it can be certain that every node in the network has knowledge of it.

We therefore introduce an acknowledgment system. For each special action generated in the network, we create a tree, rooted at the initiator node, to which nodes are added as they are first informed of the special action. When the tree is complete, acknowledgments are sent, via the tree, to the initiator node, who then sends a message to erase all memory of this special action.

Unfortunately, this modification introduces a new difficulty; failure of a node or link of a tree can disrupt the flow of acknowledgments, necessitating some type of emergency action. As in the previous algorithm, the Global Reinitialization Procedure is one possible solution to this problem, the advantage here being that only certain failures necessitate this emergency action, whereas all new links required this action previously.

CONCLUSION

We have developed a static algorithm and two dynamic algorithms that require only local information transfer and local topological knowledge. These algorithms have finite-time asynchronous convergence properties. Each has certain advantages and disadvantages, and the applicability and suitability of each is a function of the characteristics of the particular network under consideration.

ACKNOWLEDGMENT

This work was performed at Washington University, St. Louis, Missouri, and at the University of California, Santa Barbara, and was supported by the Office of Naval Research under Contracts N00014-76-C-0667 and N00014-79-C-0549.

REFERENCES

[1] J. M. Abram and I. B. Rhodes, "A Decentralized Shortest Path Algorithm," Proceedings of the Sixteenth Allerton Conference on Communications, Control and Computing, University of Illinois, Oct. 1978, pp. 271-277.

[2] T. C. Hu, Integer Programming and Network Flows, Addison-Wesley, Reading, MA, 1969, pp. 151-161.

[3] R. W. Floyd, "Algorithm 97, Shortest Path," Commun. ACM, Vol. 5, 1962, p. 345.

[4] L. R. Ford, Jr., and D. R. Fulkerson, Flows in Networks, Princeton University Press, Princeton, NJ, 1962, pp. 130-133.

[5] R. Lau, R. C. M. Persiano, and P. P. Varaiya, "Decentralized Information and Control: A Network Flow Example," IEEE Trans. Automat. Contr., Vol. AC-17, Aug. 1972, pp. 466-473.

[6] P. M. Merlin and A. Segall, "A Failsafe Distributed Routing Protocol," IEEE Trans. on Comm., Vol. COM-27, Sept. 1979, pp. 1280-1287.

Reprint of Paper:

"Some Quantitative Measures of Controllability and Observability and their Implications", Ian B. Rhodes, Proceedings of the Eighth Triennial World Congress of the International Federation of Automatic Control, Kyoto, Japan, August 24-26, 1981

SOME QUANTITATIVE MEASURES OF CONTROLLABILITY AND OBSERVABILITY AND THEIR IMPLICATIONS

Ian B. Rhodes

Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, California, U.S.A.

Abstract. Several inter-related quantitative measures of controllability, observability, reachability, and reconstructibility are introduced, their properties discussed, and some of the implications indicated. The measures vary continuously with the system parameters and the system state, and they reflect appropriate duality relationships. Two areas of application are considered: one is in quantifying the degree of interaction between system input and output, and the other is in providing simple, design-oriented upper and lower bounds on estimation performance.

Keywords. Controllability; observability; linear systems; control theory; system theory; filtering; Kalman filter; interaction; performance bounds.

INTRODUCTION

The large and highly-developed body of knowledge concerning the structural properties of linear systems is framed almost entirely in terms of "yes" and "no" questions and answers; a state is either reachable or it is not, an input disturbance is either localized away from an output or it is not, a system is decoupled or it is not; in each case, available characterizations afford conditions that can be checked to determine which of the two holds true or can be made to hold true with, for example, suitable state feedback. Almost all of these involve, directly or indirectly, the controllability and observability properties of the system. There is, however, no body of knowledge relating to the approximate achievement of these goals, or, more generally, of the degree to which they are achieved. For many practical purposes it is sufficient if, for example, an input has an acceptably small influence on an output, and it is not necessary for this influence to be zero. Especially in large systems, some measure of the degree of interaction or noninteraction between subsystems seems essential for analyzing the system, for designing estimation or compensation schemes, and for assessing the performance of these estimators and controllers (in terms, say, of performance bounds).

We introduce here several inter-related measures of reachability and observability (also controllability and reconstructibility) that are quantitative in nature and depend continuously on the system parameters and the state concerned. These measures exhibit appropriate duality relationships. We also illustrate the applications of these in two areas: The first is in quantifying the degree of interaction between system input and output, and the second is in providing upper and lower bounds on estimation performance.

For simplicity of presentation we concentrate here on constant, continuous-time systems. Entirely analogous results apply for discrete-time systems, and the extension to time-varying systems is straightforward. Extension can also be made to infinite-dimensional systems such as those satisfying partial differential equations or delay-differential equations.

MEASURES OF REACHABILITY, CONTROLLABILITY, OBSERVABILITY, AND RECONSTRUCTIBILITY

Consider the continuous-time linear system

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad (1a)$$

$$y(t) = Cx(t) \qquad (1b)$$

with $x(t)\varepsilon R^n$, $u(t)\varepsilon R^p$, and $y(t)\varepsilon R^m$.

One natural way to measure the reachability of a given state x at time T is as the maximum value of the inner product z'x over all states z to which the system (1a) can be driven at time T (from the origin at

This research was supported by the Office of Naval Research under Grant N00014-79-C-0549.

time zero) with at most 1 unit of input energy, i.e.,

$$r(x) = \max \left\{ x'z : z = \int_0^1 e^{A(T-t)}Bu(t)dt, \right.$$

$$\left. \left[\int_0^T u'(t)u(t)dt \le 1 \right] \right\}$$

Performing this straightforward maximization we find

$$r(x) = (x'W_R(0,T)x)^{\frac{1}{2}} \tag{2}$$

where $W_R(0,T)$ is the reachability Gramian (Brockett, 1970)

$$W_R(0,T) = \int_0^T e^{A(T-t)}BB'e^{A'(T-t)}dt \tag{3}$$

If x has norm 1, r(x) reduces to the projection along x of all states of (1a) that are reachable with at most one unit of input energy.

Iff r(x) = 0 then x is unreachable in the standard sense of the term, i.e., orthogonal to the set of reachable states, so the inner product x'z is identically zero; equivalently, x is in the nullspace of $W_R(0,T)$. Large values of r(x) correspond in some sense to states that are easily reached, though not in the classic sense of the term, since such states may have components that lie in the nullspace of $W_R(0,T)$ and so cannot be reached.

It can be easily shown that r(·) varies continuously with both the system parameters A and B, the final time T, and the state x.

It is also easily seen that r(x) is a sublinear function of x. It is also convex, so that, in particular, the set of states whose reachability measure is no greater than α, i.e., {x: r(x) ≤ α}, is a convex set for all α ≥ 0.

The corresponding dual observability measure of a state x at time 0 is simply the L_2-norm of the output function y over [0,T] that is produced by x, i.e.,

$$o(x) = \|y\| = (x'M_0(0,T)x)^{\frac{1}{2}} \tag{4}$$

where $M_0(0,T)$ is the observability Gramian

$$M_0(0,T) = \int_0^T e^{A't}C'Ce^{At}dt \tag{5}$$

We note that o(x) is zero iff the state x is unobservable in the standard sense of the term, i.e., x gives rise to an output that is identically zero over [0,T]; equivalently, x lies in the nullspace of $M_0(0,T)$. Large values of o(x) mean that x gives rise to an output function y with large norm, and in this sense x is highly observable.

The observability measure o(·) varies continuously with the system parameters A and C, the final time T, and the state x. It is sublinear and convex; the set of all states whose observability measure is less than a given β is convex; i.e., {x: o(x) ≤ β} is convex.

These measures, r(x) and o(x), preserve suitable duality conditions. It is easily checked that the reachability measure, $r^d(x)$, of the system dual to (1) is simply the observability measure o(x) of (1). Conversely, the observability measure, $o^d(x)$, of the system dual to (1) is the reachability measure of (1).

It is convenient, in order to avoid the square roots in (2) and (4), to work with the measures

$$R(x) = \frac{1}{2} r^2(x) = \frac{1}{2} x'W_R(0,T)x \tag{6}$$

$$O(x) = \frac{1}{2} o^2(x) = \frac{1}{2} x'M_0(0,T)x \tag{7}$$

The physical interpretations of these measures are just as one half of the squares of the quantities involved in defining r(x) and O(x): thus, R(x) is one-half of the maximum value of $(z'x)^2$ over all appropriate z, while O(x) is one-half of the square of the L_2-norm of the output function y over [0,T] that is produced by x.

Another reachability measure is the conjugate functional of R. The conjugate functional of a convex functional R is defined by (see, e.g., Luenberger, 1969),

$$R^*(z) = \sup_x [z'x - R(x)] \tag{8}$$

Performing the indicated maximization with R(x) given by (6) we find

$$R^*(z) = \frac{1}{2} z'W_R^{-1}(0,T)z \tag{9}$$

assuming that $W_R(0,T)$ is invertible; if it is not, $R^*(z)$ is still defined by (8) and takes the value ∞ if z lies in the nullspace of $W_R(0,T)$. For simplicity of notation, we assume here that $W_R(0,T)$ is invertible.

This reachability measure (9) has a simple interpretation in terms of the system (1): it is one-half of the minimum amount of control energy needed to reach state z at time T from state 0 at time 0, i.e.,

$$R^*(z) = \min \left\{ \frac{1}{2} \int_0^T u'(t)u(t)dt : \right.$$

$$\left. z = \int_0^T e^{A(T-t)}Bu(t)dt \right\} \tag{10}$$

This minimum energy is known from standard least-squares linear system theory (Brockett, 1970) to be just $z'W_R^{-1}(0,T)z$, so that $R^*(z)$ of (9) is one-half this minimum energy.

Thus the measure R^* given by (9) is in fact a measure of <u>un</u>reachability: small values of R^* correspond to small values of required input energy, and thus to relatively easily reached states; large values of R^* correspond to large required input energies and thus to states that are more difficult to reach; in particular, states in the nullspace of $W_R(0,T)$ (and thus unreachable according to the "classical" definition) have $R^*(z) = \infty$. We observe that R^* is convex, as are all conjugate functionals of convex functionals.

The corresponding conjugate functional of 0 is

$$0^*(z) = \frac{1}{2}\, z'M_0^{-1}(0,T)z \qquad (11)$$

An interpretation of this in terms of the system (1) follows by considering the problem of minimizing the norm of the linear functional on the output function y that produces the projection (or, more generally, the inner product) of the initial state x_0 on the vector z. Suppose $x_0'z$ is found as the linear functional

$$x_0'z = \int_0^T w'(t)y(t)dt \ ; \qquad (12)$$

the function w of minimum L_2-norm that accomplishes this is

$$w^0(t) = Ce^{At}M_0^{-1}(0,T)z \qquad (13)$$

so that

$$\frac{1}{2}\int_0^T w^{0\prime}(t)w^0(t)dt$$

$$= \frac{1}{2}\, z'M_0^{-1}(0,T)z = 0^*(z) \qquad (14)$$

Thus 0^* gives a measure of <u>un</u>observability: Small values of $0^*(z)$ correspond to a small effort to determine $x_0'z$ and thus to a "more observable" z than do large values of $0^*(z)$. *Note that $0^*(z) = \infty$ if z is in the* nullspace of $M_0(0,T)$, corresponding to a state that is unobservable in the standard sense of the term. As with all our measures, 0^* is convex.

The measures 0^* and R^* also preserve appropriate duality relations: the observability measure 0^{*d} of the dual to system (1) is *the reachability measure R^* of system* (1), and vice versa.

As a means of making more concrete the relationship between R and R^*, observe that if $W_R(0,T)$ is diagonalized by the similarity transformation T, so that $T'W_R(0,T)T = \Lambda = \text{diag}\,\{\lambda_i\}$, then

$$R(x) = \frac{1}{2}\, x'W_R(0,T)x = \frac{1}{2}\,(T'x)'\Lambda(T'x)$$

$$= \frac{1}{2}\sum_{i=1}^{n} \lambda_i (T'x)_i^2$$

whereas

$$R^*(x) = \frac{1}{2}\, x'W_R^{-1}(0,T)x$$

$$= \frac{1}{2}\,(T'x)^1\Lambda^{-1}(T'x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i^{-1}(T'x)_i^2$$

This makes clearer the earlier observation that R is a measure of reachability, whereas R^* is a measure of unreachability.

Entirely analogous measures of controllability and reconstructibility can also be constructed, with analogous interpretations to those above.

<u>Controllability</u>:

$$C(x) = \frac{1}{2}\, x'W_C(0,T)x \qquad (15)$$

<u>Uncontrollability</u>:

$$C^*(z) = \frac{1}{2}\, z'W_C^{-1}(0,T)z \qquad (16)$$

where $W_C(0,T)$ is the controllability Gramian (Brockett, 1970)

$$W_C(0,T) = \frac{1}{2}\int_0^T e^{-At}BB'e^{-A't}dt \qquad (17)$$

Large values of C correspond to easily-*controlled states, as do small values of* C^*. For states that are uncontrollable in the classical sense, i.e., in the nullspace of $W_C(0,T)$, $C(x) = 0$, and $C^*(x) = \infty$.

<u>Reconstructibility</u>:

$$\rho(x) = \frac{1}{2}\, x'M_R(0,T)x \qquad (18)$$

<u>Unreconstructibility</u>:

$$\rho^*(z) = \frac{1}{2}\, z'M_R^{-1}(0,T)z \qquad (19)$$

where $M_R(0,T)$ is the reconstructibility Gramian

$$M_R(0,T) = \int_0^T e^{-A'(T-t)}C'Ce^{-A(T-t)}dt \qquad (20)$$

Easily-reconstructed states have large values of ρ *and small values of ρ^*. States* that are unreconstructible in the classical sense, i.e., in the nullspace of $M_R(0,T)$, have $\rho(x) = 0$ and $\rho^*(x) = \infty$.

INTERACTION AND NON-INTERACTION

We now turn to the application of the measures introduced above to quantifying the degree of interaction between an input and an output of the system (1). We emphasize that, in this context, u in (1) may be simply a part of the total input and y in (1) may be simply part of the total system output, so that we are thinking in terms of possible applications to disturbance rejection or decoupling. We adopt as a measure of interaction between input and output the number I defined by

$$I = \sup\left\{ \frac{1}{2}\int_0^T y'(t)y(t)dt: \ y(t) = Ce^{At}x_0, \right.$$

$$x_0 = \int_{-T}^{0} e^{-At}Bu(t)dt,$$

$$\left. \frac{1}{2}\int_{-T}^{0} u'(t)u(t)dt \le 1 \right\} \qquad (21)$$

i.e., I is one half of the maximum L_2-norm of the output function y over [0,T] that can be produced by an input function u over [-T,0] with one-half of the L_2-norm of u constrained to be no greater than unity. Using the definition of 0 in (7) via (4), the interpretation of R* given by (10), and the observation that $W_R(-T,0) = W_R(0,T)$ for a constant system, it is straightforward to see that

$$I = \sup\{0(x): R^*(x) \le 1\} \qquad (22)$$

One intuitive interpretation of this expression is that large interaction between input and output is a consequence of at least some states whose observability measure 0(x) is large also being reasonably reachable (in the sense that the unreachability measure R*(x) is no greater than 1). If all states having high observability, as measured by 0(x), also have low reachability (as measured by a large unreachability R*(x)), then input-output interaction will be small. That interaction between input and output should depend on the reachability and observability of the states is to be expected: the above expression for I quantifies this dependence in terms of specified measures of reachability and observability. We remark that states x for which the supremum on the right side of (22) is achieved can be interpreted as the states through which maximum input-output interaction takes place.

The corresponding dual expression is

$$I = \sup\{R(x): 0^*(x) \le 1\} \qquad (23)$$

and this has a dual interpretation to that above. This expression follows by noting that I is simply one-half of the square of the induced norm of the linear operator L mapping inputs in $L_2[-T,0]$ to outputs in $L_2[0,T]$; L can be decomposed as $L = L_2L_1$, where L_1 maps inputs over [-T,0] into states at time 0, and L_2 maps states at time 0 to outputs over [0,T]. Because (Luenberger, 1969), $\|L^*\| = \|L_1^*L_2^*\| = \|L\| = \|L_2L_1\|$, where L* is the adjoint of L and corresponds to the system dual to (1), we have using (22) and the previously-noted dualities $0^d(x) = R(x)$ and $R^{*d}(x) = 0^*(x)$,

$$I = \|L^*\|^2 = \sup\{0^d(x): R^{*d}(x) \le 1\}$$

$$= \sup\{R(x): 0^*(x) \le 1\},$$

which is (23).

BOUNDS ON ESTIMATION PERFORMANCE

Consider now the stochastic system

$$dx_t = Ax_t dt + Bdv_t \qquad (24a)$$

$$dz_t = Cx_t dt + dw_t \qquad (24b)$$

where v and w are independent normalized Wiener processes that are independent of the initial state x_0, which is Gaussian with mean $\bar{x}_0$ and covariance Σ_0. The well-known Kalman filtering equations for $\hat{x}_t \triangleq E[x_t|\mathcal{Z}_t]$, $\mathcal{Z}_t$ being the σ-algebra generated by the output process z over [0,t], are

$$d\hat{x}_t = A\hat{x}_t dt + \Sigma(t)C'[dz_t - C\hat{x}_t dt];$$

$$\hat{x}_0 = \bar{x}_0 \qquad (25)$$

where

$$\dot{\Sigma} = A\Sigma + \Sigma A' - \Sigma C'C\Sigma + BB';$$

$$\Sigma(0) = \Sigma_0 \qquad (26)$$

or, alternatively,

$$\dot{\Sigma}^{-1} = (-A')\Sigma^{-1} + \Sigma^{-1}(-A) - \Sigma^{-1}BB'\Sigma^{-1}$$

$$+ C'C; \quad \Sigma^{-1}(0) = \Sigma_0^{-1} \qquad (27)$$

A number of upper and lower bounds on Σ(T) in terms of the reachability and reconstructibility Gramians can be derived using arguments similar to those in (Sorenson, 1968). These are matrix-ordering bounds of the form $P \ge Q$, meaning that P - Q is non-negative definite. Included here, for example, is the upper bound

$$\Sigma(T) \le M_R^{-1}(0,T) + W_R(0,T) \qquad (28)$$

from which we have immediately that the variance $a'\Sigma(T)a$ of $a'\tilde{x}_T$, $\tilde{x}_T = x_T - \hat{x}_T$, is bounded above by

$$\begin{aligned}\mathrm{Var}(a'\tilde{x}_T) &\le a'M_R^{-1}(0,T)a + a'W_R(0,T)a \\ &= 2\rho^*(a) + 2R(a) \qquad (29)\end{aligned}$$

Thus $\mathrm{Var}(a'\tilde{x}_T)$ is guaranteed to be small if $\rho^*(a)$ is small (i.e., a is easily reconstructed) and R(a) is small (i.e., a is relatively unreachable from the process noise v), an intuitively reasonable result.

A lower bound on $\Sigma(T)$ can be obtained by applying the above upper bound on Σ to the equation for Σ^{-1}. In this case we obtain

$$\Sigma^{-1}(T) \le W_R^{-1}(0,T) + M_R(0,T) \qquad (30)$$

so that

$$a'\Sigma^{-1}(T)a \le 2R^*(a) + 2\rho(a) \qquad (31)$$

Thus $a'\Sigma^{-1}(T)a$ will be large if $R^*(a)$ is large (i.e., a is difficult to reach from v) and if $\rho(a)$ is large (i.e., a is highly reconstructible), which is again intuitively expected. It is also intuitively reasonable that in bounds for this, a filtering problem, reachability and reconstructibility (rather than controllability and observability) are involved. For example, one expects (and, indeed, can show) that observability is involved in smoothing problems.

These upper and lower bounds have a number of simple consequences, some of which we now outline:

Nonlinear Observations

It is shown in (Snyder and Rhodes, 1972) that if the linear observations (24b) are replaced by the nonlinear ones

$$dz_t = h(x_t)dt + dw_t \qquad (32)$$

where h is continuously differentiable, then the conditional covariance of the state given the observations is bounded below according to

$$\mathrm{Cov}[x_T|\mathcal{Z}_T] \ge \Gamma(T) \qquad (33)$$

where

$$\dot{\Gamma} = A\Gamma + \Gamma A' - \Gamma S\Gamma + BB';\ \Gamma(0) = \Sigma_0 \qquad (34a)$$

and

$$S = E\left\{\left(\frac{\partial h}{\partial x}\right)'\left(\frac{\partial h}{\partial x}\right)\right\} \qquad (34b)$$

If S is factored as C'C, the lower bound Γ is seen to satisfy the Riccati equation (26). To this, the lower bound (30) can be applied in turn to yield

$$\begin{aligned}a'(\mathrm{Cov}[x_T|\mathcal{Z}_T])^{-1}a &\le a'\Gamma^{-1}(T)a \\ &\le 2R^*(a) + 2\rho(a) \qquad (35)\end{aligned}$$

where $\rho(a)$ might be thought of as an average reconstructibility of the system with nonlinear observations (32), since it involves

$$M_R(0,T) = \int_0^T e^{-A'(T-t)}Se^{-A(T-t)}dt \qquad (36)$$

where S is given by (34b).

Noise Scaling

If h in (32) is replaced by αh (or C in (24b) is replaced by αC), corresponding to decreasing the observation noise covariance by α^2, then M_R is replaced by $\alpha^2 M_R$, ρ by $\alpha^2\rho$, and ρ^* by $\alpha^{-2}\rho^*$. Similarly, if B is replaced by βB to reflect an increase by a factor of β^2 of the input noise covariance, then W_R is replaced by $\beta^2 W_R$, R by $\beta^2 R$ and R^* by $\beta^{-2}R^*$. The bounds given above are affected simply by these noise scalings, whereas the conditional error covariance (Σ) is not.

Additional Observations

Suppose, in addition to the observations (32) or (24b), we also have

$$d\bar{z}_t = \bar{h}(x_t)dt + d\bar{w}_t \qquad (37)$$

or

$$d\bar{z}_t = \bar{C}x_t dt + d\bar{w}_t \qquad (38)$$

where $\bar{w}$ is a normalized Wiener process independent of the initial state and all other Wiener processes introduced earlier. The right side of the Riccati equation (26) then has an additional term $-\Sigma\bar{C}'\bar{C}\Sigma$, and the lower bound Γ given by (34) an additional term $-\Gamma\bar{S}\Gamma$. In either case, the solution Σ or Γ is not affected simply by the additional term. On the other hand, it is easily seen that the reconstructibility Gramian for the combined measurements z and $\bar{z}$ is simply the sum of the reconstructibility Gramians M_R and $\bar{M}_R$ corresponding to the separate observations. Thus, $\Sigma(T)$ (or $\Gamma(T)$) is bounded below according to

$$\Sigma'(T) \le W_R^{-1}(0,T) + M_R(o,T) + \bar{M}_R(0,T) \qquad (39)$$

and

$$a'\Sigma^{-1}(T)a \le 2R^*(a) + 2\rho(a) + 2\bar{\rho}(a) \qquad (40)$$

The effect of the additional observation on (here) the lower bound is extremely simple, permitting immediate assessment of the utility of additional observations or comparison of alternative observations.

Measurement Design Problems

The bounds also provide a simple framework for measurement design. For example, for a fixed vector a of interest, we can consider the maximization of the right side of (31) with respect to C. This is equivalent to maximizing

$$
\begin{aligned}
2\rho(a) &= a'M_R(0,T)a \\
&= a'\int_0^T e^{-A'(T-t)}C'Ce^{-A(T-t)}dt\ a \\
&= \text{trace}\ \ C'\int_0^T e^{-A(T-t)}aa'e^{-A(T-t)dtC}\Big\} \\
&= \text{trace}\left\{C'\bar{\bar{W}}_C(0,T)C\right\}
\end{aligned}
$$

where, (making change of variables s = T - t),

$$\bar{\bar{W}}_C(0,T) = \int_0^T e^{-As}aa'e^{-A's}ds$$

is the controllability Gramian of the system (A,a,C). If, for example, C is a row vector, the solution to this optimization problem is simply to take the vector C' along the eigenvector corresponding to the maximum eigenvalue of $\bar{\bar{W}}_C(0,T)$.

CONCLUSIONS

The quantitative measures of controllability, observability, reachability, and reconstructibility introduced here have desirable continuity and duality properties. Their applications include measuring the degree of interaction between system input and output, and providing a design-oriented framework for evaluating estimation performance in terms of upper and lower bounds.

REFERENCES

Brockett, R.W. (1970). Finite-Dimensional Linear Systems, Wiley, New York.

Luenberger, D.G. (1969). Optimization by Vector Space Methods, Wiley, New York.

Snyder, D.L., and I.B. Rhodes (1972). Filtering and control performance bounds with implications on asymptotic separation," Automatica, Vol. 8, 747-753.

Sorenson, H.W. (1968) "Controllability and observability of linear, stochastic, discrete-time control systems," in C.T. Leondes (Ed.), Advances in Control Systems, Vol.6, Academic Press, New York.

Reprint of Paper:

"Some Shortest Path Algorithms with Decentralized Information and Communication Requirements", Jeffrey M. Abram and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-27**, No. 3, pp. 570-582, June 1982.

Some Shortest Path Algorithms with Decentralized Information and Communication Requirements

JEFFREY M. ABRAM AND IAN B. RHODES, MEMBER, IEEE

Abstract—**This paper presents several decentralized algorithms for finding all shortest paths in a network. Both static and dynamic networks are discussed. Each algorithm has localized information and communication requirements, operates asynchronously, and converges to the optimum in finite time.**

I. Introduction

In recent years a major area of research has been the development of decentralized decision, estimation, and control schemes for large-scale systems. The concept of decentralization has applications to many different areas of systems theory. For example, Lau *et al.* [1] developed a decentralized algorithm for solving the problem of maximum flow through a network; minimum-delay routing has been considered by Gallager [2]; Sandberg [3] developed a decentralized scheme for synchronizing digital transmission systems; and the question of stability for large-scale systems has been addressed by Šiljak [4], among others. The reader is also referred to the April 1978 issue of the IEEE Transactions on Automatic Control, which was devoted to large-scale systems, and in particular the survey paper by Sandell *et al.* [5] and the references therein.

In this paper we are concerned with the decentralized shortest path problem and the development of algorithms that enable each node in a network to calculate its shortest distance to any other node using only local knowledge of the network topology and only local information transfer between adjacent nodes. Shortest path problems arise in many contexts, and algorithms with decentralized requirements are of obvious importance in many applications.

Our intention has been to develop algorithms general enough to be applicable in a variety of situations. We have had the same goal of generality for the network models used and the assumptions made. While our algorithms have potential application to a wide variety of specific situations, including computer-communication networks, we do not consider the specific characteristics of such networks. We also point out that in some situations the shortest path problem, which we discuss, is different from the optimal-routing problem.

The main result presented in this paper is a decentralized shortest path algorithm for dynamic networks in which branch lengths can increase or decrease, and branches or nodes can be added to or removed from the network. This algorithm operates asynchronously, has local information and communications requirements, and is guaranteed to converge to the optimum in finite time. The foundation of this algorithm is our static algorithm [6], which is based primarily on an algorithm of Ford and Fulkerson [7], and which is similar in spirit to those found in [8]–[10], and elsewhere.

The static algorithm is summarized in Section II. Section III contains the main version of the dynamic algorithm. Some alternatives to and modifications of the dynamic algorithm are given in Section IV; for some of these convergence proofs exist, while for others, which appear intuitively to lead to more rapid convergence or greater robustness, proofs have yet to be found. Finally, convergence proofs and precise descriptions of the static algorithm and the main dynamic algorithm can be found in Appendix I.

II. The Static Algorithm

Consider a directed graph consisting of N nodes, denoted $A=\{1,2,\cdots,N\}$, and a collection of links (branches) $L=\{(i,j): i,j\in A$ and there exists a link from i to $j\}$. There is at most one link from any node i to any node j, and no link from any node to itself. It is convenient, though not necessary, to assume that the existence of a link from i to j implies the existence of one from j to i. In some situations, network links are used also for communication, but it may be that network links carry certain commodities and are not used to carry algorithm-related information. For instance, in a transportation network links may carry vehicles, and a separate communication network, such as the telephone network, may be used for transmitting the information required by the algorithm. In that case, we assume that the existence of a pair of oppositely directed network links connecting two nodes implies the existence of a bidirectional communication link between them. To each link (i,j) is associated a distance $s(i,j)$, which could

Manuscript received December 15, 1980; revised June 16, 1981. Paper recommended by W. S. Levine, Chairman of the Networks and Transportation Systems Committee. This work was supported by the Office of Naval Research, Mathematical and Information Science Division under Contracts N00014-76-C-0667 and N00014-79-C-0549.

J. M. Abram was with the Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130 and the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106. His is now with Advanced Information and Decision Systems, Mountain View, CA 94040.

I. B. Rhodes is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

0018-9286/82/0600-0570$00.75 ©1982 IEEE

represent physical distance, delay, energy, money, or any other quantity appropriate to the network under consideration. The lengths $s(i, j)$ and $s(j, i)$ may, and generally do, differ. Lengths are unrestricted in sign, but to guarantee the existence of loopless shortest paths the sum of the lengths in any closed loop is assumed to be positive.

Very little topological information is needed. Each node needs to know only the identity of its neighbors, the length of the branch to, but not from, each neighbor, and the identities of all possible destinations. (Actually, a list of all destinations need not be known initially. It is possible to construct the list during the course of the algorithm.)

A. The Algorithm

The algorithm itself is rather simple. For each destination, a node calculates and stores an assessment of the shortest distance, based on information received to date, from each of its neighbors. The smallest of these is taken to be its assessment of the shortest distance to that destination, and is subsequently referred to as the *current shortest distance*. Also stored is the identity of the neighbor(s) via which this minimal distance is achieved [i.e., the next node(s) in the shortest path(s)]. Whenever a node's current shortest distance to a destination decreases, either through initialization or new information received from a neighbor, the node transmits this distance to each of its neighbors, thus allowing each neighbor to update the corresponding distance assessment by adding the distance just received to the known length of the link to the sender. Distance messages from a node to any of its neighbors are assumed to be transmitted directly along the (communication) branch connecting them, and received in the order in which they are sent. The algorithm continues in this manner until no further changes can be made. At this point, which must occur in finite time, each node knows the length of the shortest path to each destination (or that no path exists), the next node in this path, and the shortest distance via each alternative neighbor. A precise description of this algorithm appears in the Appendix.

B. Convergence of the Algorithm

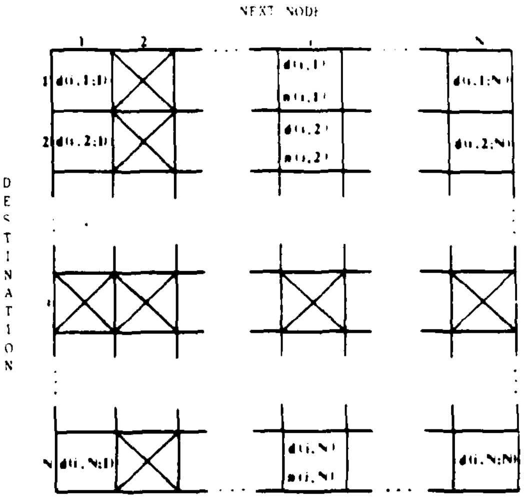


Fig. 1. Distance matrix for node i.

It is convenient for purposes of presentation to arrange the data stored by each node in an $N \times N$ matrix. Fig. 1 shows the matrix maintained by node i; each row corresponds to an *ultimate* destination, each column to an adjacent node that represents the next node in various paths to the final destinations. Row i, which would represent distances from node i to itself, and columns corresponding to nodes not adjacent to i are crossed out, except for column i, which serves a special function that is described subsequently. The quantity $d(i, k; j)$, stored in the (k, j) element, is the distance assessment from node i to node k via next node j, $d(i, k) = \min_j d(i, k; j)$ is the current shortest distance from i to k, and $n(i, k)$ the neighbor(s) achieving the distance $d(i, k)$; both $d(i, k)$ and $n(i, k)$ are stored in the kth element of column i, this being the special role of the ith column. Diagonal elements represent direct distances, i.e., lengths of links from node i to its neighbors.

In Fig. 2, the distance matrices are stacked vertically to form a distance cube. While the topological and distance information necessary to construct and update the distance cube is not localized, so that the cube is not available to any node of the network, we can utilize the concept of the distance cube in order to observe and analyze the behavior of the algorithm.

Node i's current shortest distance to final destination k, stored as $d(i, k)$ in the (k, i) element of i's distance matrix, is also reflected in the (k, i) element of each of node i's neighbors; in particular, the (k, i) entry in node j's distance matrix is $s(j, i) + d(i, k)$. Transmission of any change in $d(i, k)$ must therefore be reflected in the distance cube by information transfer along a vertical line through the (k, i) element of each node's distance matrix. Thus, communication of distance information occurs along vertical lines in the cube. Also, in a given node's distance matrix, the operations in any row are self-contained and independent of those in any other row; the initialization step and each subsequent operation of updating a distance assessment and comparing it to the current shortest distance involves only elements in a specific row. Together, these observations reflect the fact that the "vertical slice" corresponding to each fixed destination is self-contained insofar as both communication and addition-comparison operations are concerned. Convergence of the algorithm is most easily proven by decomposing the distance cube into vertical slices for which convergence can be proven separately and individually. The following is an outline of a straightforward proof for the static algorithm that can be found in [6]. The Appendix contains a detailed, but somewhat different, proof that is better suited to the subsequent proof for the dynamic algorithm.

Consider the vertical slice corresponding to any destination, say node k. Any node whose direct path to k is a shortest path, has this distance available initially. Any other distance assessment for destination k possessed by the node must be greater than or equal to this shortest

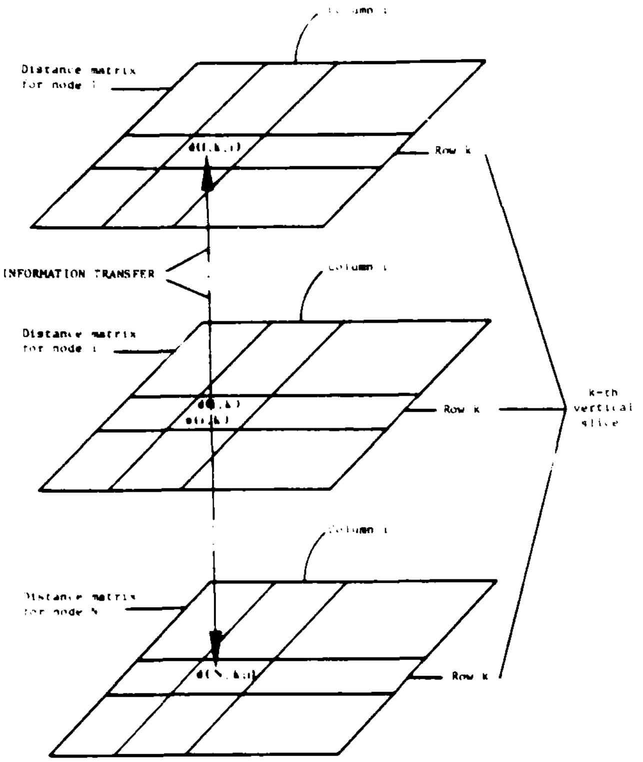


Fig. 2. Distance cube.

distance. Although the node is unaware that the direct distance to k is optimal, it must find the distance to be the minimal element in row k, take it to be the current shortest distance to node k, and transmit this distance to each of its neighbors. Similarly, a node that has a shortest path containing two links constructs the shortest distance to k after receiving the optimal distance from the intermediate node in the path (which, by the principle of optimality, must be a node whose direct path to k is a shortest one), and takes that distance, again unaware of its optimality, to be its current shortest distance. In this manner, knowledge of shortest distances spreads until each node in the network has found the shortest distance to k. The same is true for each destination, and every distance in the distance cube is guaranteed to converge to the optimum in finite time.

We note that the convergence proof focuses only on the construction and transmission of *optimal* distances, ignoring the suboptimal distances which may be transmitted prior to the construction of shortest distances. This focusing on optimal information brings out a resemblance between the static algorithm and dynamic programming. For a given destination, think of the destination as the final dynamic programming stage. The set of nodes that have a shortest path containing one link comprises the next-to-last stage. The stage preceding this consists of all nodes having shortest paths with exactly two links, and so on. (Note that a node can be in several different stages if it has more than one shortest path.) When the nodes in stage n have found their shortest distances to the destination and have sent them to their neighbors, the nodes in stage $n-1$ are then able to find their shortest distances. Thus, the static algo-

programming works. This dynamic programming comparison is appropriate to the analysis, but not the actual operation, of the static algorithm.

C. Convergence Time

Since asynchronous operation of the static algorithm allows a node to wait indefinitely before transmitting new information, convergence time can be bounded only by bounding this delay. Suppose δt is the maximum amount of time required for a node to receive, process, and relay a distance message. Using the convergence proof, it can be seen easily from the discussion above that a node having a shortest path containing the maximum possible $N-1$ links must possess the corresponding shortest distance after at most $(N-2)\delta t$ units of time, and this is therefore an upper bound on the time taken for each node to know its shortest distance to every other node. After one more unit of time, every node knows "everything," i.e., each element of the distance cube achieves the optimum in at most $(N-1)\delta t$ units of time. Bounding the number of *operations* is much more difficult. A more detailed discussion of convergence bounds can be found in [11].

III. Dynamic Networks

While the above algorithm handles static networks, many problems arise for a dynamic network in which branch lengths may decrease or increase, new branches (nodes) may be introduced into the network, and branches (nodes) may be removed from the network. It should be noted that adding or removing a node can be treated by simultaneously adding or removing the incident set of branches. Again, we point out that there may be a distinction between network links and communication links. If so, we assume that a network link can become infinite in length without being removed from the network, provided that the corresponding communication link remains intact. A network link can be removed from the network, in which case the communication link is also removed. Or when adding new links to the network, a communication link can be activated before the corresponding network link. But we assume that a communication link cannot be removed if the corresponding network link remains functional.

The easiest case to handle is that of decreasing branch lengths. The algorithm is easily modified as follows: a node detecting a decreased branch length to some neighbor decreases all distance assessments via that neighbor by the amount of the detected change, then performs a comparison operation in each affected row to find the new current shortest distance to the ultimate destination represented by that row. The proof is very similar to the one previously outlined. Suppose that a finite number of branch length decreases occur before some time t_f, after which no decreases occur for some period of time. The topology at time t_f determines the set of shortest distances to which the algorithm should converge. A crucial element of the convergence proof for the static algorithm is that, at any time, each distance assessment is no smaller than the correspond-

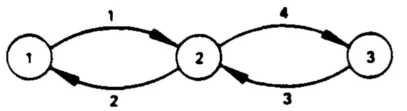


Fig. 3. Dynamic network for which static algorithm does not converge.

appears in the distance matrix, it is the smallest element of its row and is accepted as the current shortest distance. Since branch length decreases can cause decreases, but not increases, in actual shortest distances, this condition is met for the shortest distances at time t_f, and convergence to these values follows from the original static convergence proof.

On the other hand, an increased branch length can cause an increased actual shortest distance, violating this condition and invalidating the original convergence proof. In this case, the algorithm may not converge to the optimal solution in finite time, as illustrated by the example in Fig. 3.

For this example, we restrict attention to destination 3, and assume that nodes 1 and 2 have converged to the obvious shortest distances of 4 and 5, respectively. Additionally, node 2 has optimal distance of 7 via node 1. Suppose that branch length $s(2,3)$ increases to infinity. Node 2, observing this increase, would change its current shortest distance to the distance of 7 via node 1, and transmit this new data to 1. (Note that the *actual* shortest distance of infinity is possessed by node 2, but the distance assessment 7 is smaller than this shortest distance.) The current shortest distance for node 1 then becomes 8, and is sent to node 2. This increases the shortest distance for node 2 to 10, which is then sent to node 1, etc. The distance assessments to node 3 for both nodes 1 and 2 increase without bound, but do not reach the correct value of infinity in finite time.

One could remedy this problem by establishing a maximum possible shortest distance. Any distance exceeding this maximum could be taken as infinity. Nevertheless, the convergence time of the static algorithm can be very large when increases occur, as in the example above.

This example illustrate the fact that a single topological change can prevent convergence of our static algorithm. Even for the dynamic algorithms that we have developed to accommodate topological changes, convergence is not possible if these changes occur with such frequency that the optimal solution always changes before the algorithm has time to converge to a solution. Therefore, we must assume for each of our dynamic algorithms that topological changes cease for some period of time, sufficiently long to allow convergence to take place, after which topological changes may resume. A sufficient period without topological change would be $2(N-1)\delta t$, where δt is as defined in the previous section.

We point out that our definition of convergence may differ from that used elsewhere. For instance, suppose each distance in a network oscillates between two values as a function of the routing scheme being used, as in an example presented by Bertsekas [13] in the context of computer-communication networks. Though the set of shortest paths oscillates between two solutions, we consider any algorithm that can find these shortest paths to be convergent.

Since the performance of the static algorithm is poor for dynamic networks, we have developed several modified versions of the static algorithm, some of which are described in this paper. Each of these algorithms is based primarily on some form of reinitialization of the basic static algorithm; they differ mainly in the mechanics of the reinitialization. The remainder of this section is devoted to our main dynamic algorithm. The algorithm contains a procedure to treat branch length increases and removal of branches and a procedure for adding new links.

A. Assumptions

For the dynamic case, we make the following assumptions about the network and the performance of the individual nodes.

Assumption 1: There exist times t' and t_f such that any finite number of topological changes can occur between t' and t_f, but that none occurs for some period of time after t_f.

Assumption 2: Each node knows the distance to each of its neighbors at all times.

Assumption 3: At t', the distance assessment from i to k via j is greater than or equal to the sum of the length of link (i, j) and the current shortest distance $d(j, k)$ for any choice of i, k, and j. That is, node i cannot have knowledge of a better distance from j to k than node j has.

Assumption 4: At t', each distance assessment $d(i, k; j)$ is no smaller than the actual shortest distance from i to k.

Assumption 5: There exists an upper bound Δt, known to each node in the network, on the amount of time it can take a message to traverse any loopless path in the network, then traverse one more link. A possible choice for Δt would be N times the maximum possible processing and transmission time required for a message to be sent between neighbors. Note that calculating Δt may require some centralized information, although a decentralized scheme for determining Δt could probably be developed.

Assumption 6: Each node is equipped to measure elapse of time, although a time base common to all nodes is not required.

B. The Local Reinitialization Procedure

It is not sufficient to merely disseminate a reinitialization command throughout the network when a branch length increase or removal of a branch occurs. Once a node has reinitialized, it needs some guarantee that subsequently received information is based on all affected nodes having reinitialized also. Thus, a mechanism has been devised for each node to determine which distance information it receives is trustworthy (in that all nodes further down the corresponding path are guaranteed to have reinitialized) and which is not. In simple terms, on hearing that a branch length increase or branch removal has taken place, a node ignores distance information (but not other messages) sent by any questionable neighbor until that neighbor acknowledges that it, too, is aware of the change. As each neighbor in turn so acknowledges, the embargo on its information is removed. In this way, some convergence toward the new solution can be taking place while news of the change is still propagating through the network.

More precisely, a *local reinitialization message* is initiated by a node detecting an increase in length or removal of any link. When this occurs, the initiator assigns a unique index to the generated message (consisting of the initiator's identity and a counter), and does the following:

Step 1: Reinitializes by eliminating every distance assessment in memory except for direct distances, and by determining new current shortest distances.

Step 2: Places an embargo, with the proper index, on distances received from each of its neighbors.

Step 3: Transmits all of its current shortest distances to each of its neighbors, along with the indexed local reinitialization message.

A node other than the initiator takes similar action upon receiving a particular local reinitialization message for the first time. Each embargo placed by a node due to a local reinitialization is removed as the local reinitialization message is received from the corresponding embargoed neighbor. Specifically, when a message is received from a neighbor, a node's action is governed by the following:

Case 1: Message contains no local reinitialization.

a) If there is no embargo on this neighbor, calculate the new distance assessments via this neighbor, perform addition-comparison operations, and transmit any new current shortest distances to each neighbor.

b) Otherwise, ignore the message.

Case 2: Message contains a local reinitialization.

a) If there is no embargo with matching index already on this neighbor, calculate the new distance assessments via this neighbor (unless another embargo is on the neighbor, in which case distance calculations are omitted), and perform Steps 1–3 above, except that the distance assessments from the neighbor just heard from are not deleted or embargoed.

b) If there is a matching index, remove that embargo from this neighbor. If no other embargo exists on the neighbor, take action as in Case 1a). Otherwise, ignore the distance component of the message and take no further action.

Any finite number of local reinitializations can exist within the network simultaneously, each distinguishable by its index. A branch removal is treated by eliminating the appropriate node from the set of neighbors.

C. The Procedure for Adding Links

The above procedure alone would be sufficient if the addition of new links and nodes did not occur. The crucial goal of the procedure is to guarantee to each node that any neighbor affected by an increase has reinitialized before distance information from it is accepted. If a node sends a local reinitialization message to each of its neighbors, then acquires a new neighbor through the activation of a new link, there would be no guarantee that this new neighbor had recently reinitialized or that its distance information

To eliminate this problem, suppose that two nonadjacent nodes wish to activate a new pair of oppositely directed links between them, and that they make initial contact for this purpose at the time t_{new}. (If communication links differ from network links, this contact is made possible by activating the communication link before the network link is added. In any event, we assume that some means of communication exists in this situation.) Then, in order to activate these links, each node executes the following steps:

Step 1: During the interval $[t_{\text{new}}, t_{\text{new}} + \Delta t)$, where Δt is the maximum possible transmission delay between any pair of nodes, as defined in Section II-A, store the index of each local reinitialization message received or initiated. Also, in consultation with the tentative new neighbor, determine which of the two will initiate a local reinitialization message when the links are activated.

Step 2: At time $t_{\text{new}} + \Delta t$, activate the new links, and add the new neighbor to the set of neighbors. Then, if this node is the initiator agreed upon in Step 1, initiate a local reinitialization message.

Step 3: During the interval $[t_{\text{new}} + \Delta t, t_{\text{new}} + 2\Delta t)$, compare the index of any local reinitialization message received from the newly acquired neighbor to the indexes stored in Step 1. If this index is in the set, do not reinitialize. Rather, transmit the local reinitialization back to the new neighbor, then proceed as in Case 1 of the previous section. If this index is not in the set, proceed with the normal reinitialization steps.

Step 4: At time $t_{\text{new}} + 2\Delta t$, remove the indexes stored in Step 1 from memory, completing the new-link procedure.

The waiting period of Δt and the storing of indexes received during this period serve to prevent a particular local reinitialization message from causing any node to reinitialize more than once, a helpful fact in proving convergence. The validity of distance information exchanged by the two new neighbors is guaranteed by the local reinitialization message initiated upon activation of the new links.

The two procedures just described, one for initiating local reinitialization messages and one for adding new links, are the major components of our main dynamic algorithm. This algorithm is described in detail in the Appendix.

D. Convergence

The most important property of the dynamic algorithm is the guarantee, given our assumptions, that it will converge to the optimal solution of the shortest path problem, regardless of the number and type of topological changes that occur. The convergence proof, which appears in detail in the Appendix, is outlined here.

After the time t_f, defined in Assumption 1, the topology of the network is fixed for some period of time, and consists of one or more components, i.e., maximal connected subnetworks. Consider any component. If no node in this component has ever initiated a local reinitialization message, then the algorithm behaves like the static one with respect to this component. Otherwise, there exists a local reinitialization message that was the last one prior to *t*, initiated within the component. The component must be onnected at this initiation time, guaranteeing that each ode of the component receives the message and reinitializes after the final initiation time. Since no further increases or failures can occur in the component (since it is ssumed that no further local reinitializations are generated), and since reinitialization is guaranteed for each component node at a time after which only decreases in istances can occur, we can prove the existence of a time fter which each distance assessment in the component is no smaller than the corresponding shortest distance. The ıme is true for each component.

Then it is shown that once an embargo is placed on a neighbor, it must be removed either through a reply from hat neighbor or removal of the neighbor from the set of eighbors. Also proven is that once an embargo is lifted, it cannot be replaced. Thus, there exists a time after which all embargoes in the network have been removed. This reıoval of all embargoes guarantees to each node the availaıility of a distance assessment via each neighbor.

Finally, once all embargoes have been removed and istance assessments are no smaller than corresponding hortest distances, the dynamic algorithm behaves like the static one, and convergence to the optimal solution follows om the guaranteed convergence of the static algorithm.

Convergence for this dynamic algorithm is guaranteed to occur within $2(N-1)\delta t$ units of time after the "final" opological change; it can take $(N-1)\delta t$ units for each ode to have reinitialized and an additional $(N-1)\delta t$ units for convergence to the optimum for all possible destinations. One would expect average convergence time to be far iperior to this worst case bound.

The dynamic algorithm requires more memory than the static one in order to keep track of embargoes and indexes. requires that nodes activating new links have clocks and ıowledge of Δt, determination and dissemination of which may require some centralized information, although Δt ould probably be determined in a decentralized fashion. ut aside from the knowledge of Δt, the algorithm is informationally decentralized, requires only local comıunication, is implemented asynchronously, and is ıaranteed to converge to optimality.

IV. Alternative Algorithms

Each of the following alternative algorithms requires, although each for a different purpose, occasional use of hat we refer to as a *global reinitialization procedure*. This rocedure brings about global reinitialization by effectively suspending all communication of distance information for sufficiently long period of time (i.e., Δt) to ensure that all odes have reinitialized. Several possible mechanisms for achieving this present themselves: one is for the initiator of the action to decide upon a future time at which communiıtion of distance information based on reinitialized data ill resume, and to send this time to each of its neighbors, which continue to propagate the message throughout the etwork. Implicit here is the existence of a time base ommon to all nodes, and the availability to each node of the value of Δt, which may or may not require some global information, as mentioned earlier. (Other mechanisms exist for which a common time base is unnecessary.)

While the message to reinitialize is received by different nodes at different times, there is a simultaneous resumption of communication of distance information, hence the reinitialization is effectively a *global* one. This differs greatly from the *local* reinitialization of the previous section. There, a node reinitializes upon receipt of a message to do so, and begins transmitting distance information immediately thereafter.

Perhaps the most practical utilization of the global reinitialization procedure is as a replacement for the new-link procedure in the previous dynamic algorithm. The procedure is initiated by either of the nodes wishing to bring up the new link. By empowering the global reinitialization to eliminate embargoes, when communications resume at the agreed upon time each node receiving the global reinitialization message has removed all embargoes, and has no distance assessments except for direct distances. Some nodes will not receive this message if the network is not connected, but they will be unaffected by the new link. Thus, the new link is added without allowing an old local reinitialization message to be transmitted over it, preventing a second reinitialization by a node that received it previously, and preventing incorrect distance information from being exchanged by the new neighbors. This alternative algorithm can be proven to converge to optimality.

Another alternative, also proven to converge, involves superimposing an acknowledgment system on the local reinitialization procedure, which is then used to simplify the introduction of new links. In this algorithm, when a node, say i, first receives any particular local reinitialization, it remembers the identity of the node sending the message, that node being referred to as the *upstream neighbor* for this reinitialization. Each node that first receives the message from node i is a *downstream neighbor* of i, and a tree is created as the message spreads downstream. (Actually, the same type of tree is implicitly generated in the dynamic algorithm of Section III.) When a node has no downstream neighbor, either because it has no neighbors besides the upstream one or because all of its other neighbors have already joined the tree, or when a node has received a *special acknowledgment* from each of its neighbors, it then sends special acknowledgment to its upstream neighbor. Embargoes are placed and removed as before, but the index of each local reinitialization is remembered until a node is told otherwise. This occurs when the initiator has received a special acknowledgment from each of its neighbors, reflecting the fact that every node connected to the initiator by some path has received the local reinitialization. The initiator then generates a message that is propagated through the network to erase the tree and all memory of the corresponding index.

The new-link mechanism for this algorithm is very similar to the one presented earlier, except that instead of remembering indexes only during the interval $[t_{\text{new}}, t_{\text{new}} + \Delta t)$, the necessary indexes are already in memory at t_{new}, thus eliminating the delay of Δt when activating a new link. As before, a local reinitialization is generated when the link is activated.

The global reinitialization procedure is used in this algorithm as an emergency action. A removal of a link or node of a tree can disrupt the flow of acknowledgments to the initiator, or of "erasure" messages from the initiator. To guarantee the eventual removal of all memory of indexes in the event of removal of a tree branch, the global reinitialization procedure is invoked. Without this action, convergence would still occur, but some indexes may be remembered indefinitely, long after they are no longer needed.

In addition to the disadvantage of requiring the use of the global reinitialization procedure in this algorithm, additional memory is required for each node to remember all indexes received, when actually indexes need be remembered only when adding a new link and only by the two nodes adding the link.

On the other hand, it is an algorithm that is guaranteed to converge to optimality in a dynamic network. Also, we feel that the concept of the acknowledgment system has potential value in other network situations where a guarantee that every node has received some message may be required.

This alternative algorithm is similar to one developed by Merlin and Segall [12] in that both utilize acknowledgment systems, but many differences between the two algorithms exist. A major difference is that in our algorithm the node detecting the topological change generates the reinitialization; in theirs each destination controls the updating of paths. To solve the shortest path problem for all destinations, the Merlin/Segall algorithm would apparently require that messages requesting update be sent to all possible destinations after each critical topological change.

The upper bound on convergence time for each of the alternative algorithms is $2(N-1)\delta t$ after the "final" change, as was the case for the main dynamic algorithm. An additional $(N-1)\delta t$ may be required in the acknowledgment algorithm before all trees are removed.

Several modifications have been developed for the algorithms presented thus far. Information storage requirements can be reduced by storing current shortest distances, but not distance assessments. The assumption that network links are bidirectional can be relaxed and the algorithms modified accordingly. The number of reinitializations generated can possibly be reduced significantly by requiring a node to generate such action only when a link length increase satisfying certain conditions occurs. Alternative procedures for adding new links have also been developed. These modifications and others are discussed in [11].

V. Conclusions

Our goal in this paper has been to present an application of decentralized information and control to the solution of the shortest path problem. The algorithms that we have developed to solve this problem share several desirable properties.

Each algorithm operates asynchronously; a node acts only when the situation warrants action. Moreover, each algorithm can be implemented without even the need for a time base common to every node in the network. Different nodes can have different capacities for processing and transmitting algorithm-related messages. These messages can be acted upon and subsequently transmitted by the various nodes in any order.

The communication requirements of each algorithm are totally decentralized, the topological information requirements decentralized except for knowledge of Δt; Δt can probably be determined in a decentralized fashion, a possibility that is currently under investigation. The highly localized nature of the algorithms limits their computational efficiency. Efficiency can be improved by giving nodes access to additional information, as in the algorithm in [14], for instance, in which each node is allowed to construct the entire network topology. Perhaps this degree of efficiency can be matched or exceeded by algorithms requiring only some centralized information.

The most important property shared by our algorithms is that each can be proven to converge to the optimum in finite time. This contrasts with most other applications of decentralized control in which suboptimal performance is the cost of decentralization.

Appendix
Notation

The following notation and definitions are used in the proofs and flowcharts presented in this Appendix:

$d_t(i,k;j)$	The distance assessment from i to k via j at time t.
$d_t(i,k)=\min_j d_t(i,k;j)$	The current shortest distance from i to k at time t.
$n_t(i,k)$	The next node(s) achieving the distance $d_t(i,k)$.
$s_t(i,j)$	The length of link (i,j) at time t.
$d_t^*(i,k)$	The actual shortest distance from i to k at time t.
$d_f^*(i,k;j)$	The final shortest distance from i to k via j.
$d_f^*(i,k)$	The final shortest distance from i to k.
$s_f(i,j)$	The final length of link (i,j).
Δt	An upper bound on the maximum possible time for a message to traverse any loopless path plus one more link (which may form a loop).
COUNTER (i)	The index counter last used by node i in generating a local reinitialization message.

Convergence of the Static Algorithm

The following convergence proof for the static algorithm given in Fig. 4 is designed to facilitiate its application to proving convergence for the dynamic algorithm. A detailed version of the simpler static convergence proof that is

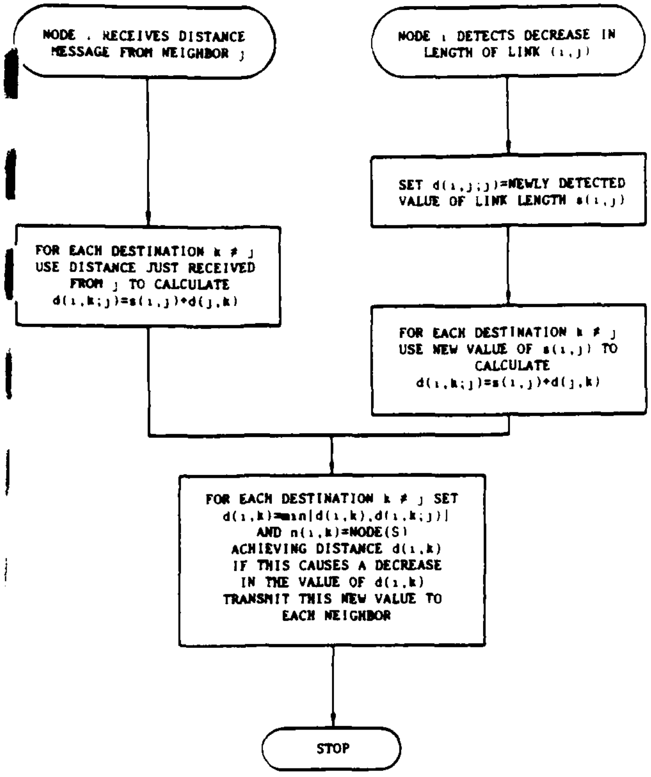

NODE ı RECEIVES DISTANCE MESSAGE FROM NEIGHBOR j
NODE ı DETECTS DECREASE IN LENGTH OF LINK (ı,j)
SET d(ı,j;j)=NEWLY DETECTED VALUE OF LINK LENGTH s(ı,j)
FOR EACH DESTINATION k ≠ j USE DISTANCE JUST RECEIVED FROM j TO CALCULATE d(ı,k;j)=s(ı,j)+d(j,k)
FOR EACH DESTINATION k ≠ j USE NEW VALUE OF s(ı,j) TO CALCULATE d(ı,k;j)=s(ı,j)+d(j,k)
FOR EACH DESTINATION k ≠ j SET d(ı,k)=min[d(ı,k),d(ı,k;j)] AND n(ı,k)=NODE(S) ACHIEVING DISTANCE d(ı,k) IF THIS CAUSES A DECREASE IN THE VALUE OF d(ı,k) TRANSMIT THIS NEW VALUE TO EACH NEIGHBOR
STOP

Fig. 4. Actions taken by node i during static algorithm.

utlined in Section II of this paper can be found in [6].

Theorem 1: Suppose that the static algorithm is used at ·ll times after some time T, regardless of what may have een used before T, and that the following conditions are satisfied.

Condition 1: No topological or distance changes occur in ıe network after time T.

Condition 2: $d_t(i, k; k) = s_t(i, k)$ for any $t \geqslant T$, and for any i and k.

Condition 3: $d_t(i, k; j) \geqslant s_t(i, j) + d_t(j, k)$ for any $t \geqslant T$, nd for any i, k, and j, $k \neq j$. (If $d_t(i, k; j)$ is acquired by node i by some means other than adding observed length (i, j) to the distance $d_t(j, k)$ received from node j, then ıis condition guarantees that $d_t(i, k; j)$ cannot increase as the result of information received from node j. In other words, node i cannot have access to better information on distance via node j than node j itself has. This condition is required to guarantee that $d_t(i, k) = \min_j d_t(i, k; j)$ in the convergence proof.)

Condition 4: $d_t(i, k; j) \geqslant d_f^*(i, k)$ for any $t \geqslant T$, and for any i, k, and j, and therefore $d_t(i, k) \geqslant d_f^*(i, k)$, also.

Condition 5: Consider any $d_T(i, k)$. It is assumed that ıode i has sent or will send this distance to each node that s a neighbor of node i in the final topology. Furthermore, if the final transmission of this distance by node i is eceived by any neighbor j prior to time T, then $d_T(j, k; i)$ $= s_f(j, i) + d_T(i, k)$.

Then, there exists a time T^* after which each entry in the istance cube, i.e., each current shortest distance and each istance assessment, is optimal with respect to the final ›pology.

Proof: Proof is accomplished by proving a series of lemmas. The conditions for Theorem 1 are assumed to hold for each of these lemmas.

Lemma 1: For any $d(i, k)$, there exists a time t_{ik} such that

$$d_{t_{ik}}(i, k) = d_f^*(i, k). \tag{1}$$

Proof: If $d_T(i, k) = d_f^*(i, k)$, then (1) is true for $t_{ik} = T$. Otherwise, by Condition 4,

$$d_T(i, k) > d_f^*(i, k). \tag{2}$$

This implies that $d_f^*(i, k)$ is finite, which implies the existence of at least one path from i to k. Since there are a finite number of paths from i to k in the network, there must be a shortest path. Suppose one such shortest path is

$$i = m_0 \rightarrow m_1 \rightarrow m_2 \rightarrow \cdots \rightarrow m_p = k. \tag{3}$$

By the principle of optimality, for any m_r in (3),

$$m_r \rightarrow m_{r+1} \rightarrow \cdots \rightarrow m_p = k$$

is a shortest path from m_r to k. In particular, the direct path from m_{p-1} to k is a shortest path, and

$$d_f^*(m_{p-1}, k) \leqslant d_T(m_{p-1}, k) = \min_j d_T(m_{p-1}, k; j)$$
$$\leqslant d_T(m_{p-1}, k; k) = s_T(m_{p-1}, k) = d_f^*(m_{p-1}, k).$$

Thus,

$$d_T(m_{p-1}, k) = d_f^*(m_{p-1}, k). \tag{4}$$

By (2), (4), and the optimality of (3), there exists a largest index l, $l < p-1$, such that $d_T(m_l, k) > d_f^*(m_l, k)$. Since l is maximal, $d_T(m_{l+1}) = d_f^*(m_{l+1}, k)$. The suboptimality of $d_T(m_l, k)$ and Condition 5 imply that node m_l will receive the distance $d_f^*(m_{l+1}, k)$ from node m_{l+1} at some time $T_l > T$, at which

$$d_f^*(m_l, k) \leqslant d_{T_l}(m_l, k) \leqslant d_{T_l}(m_l, k; m_{l+1})$$
$$= s_f(m_l, m_{l+1}) + d_f^*(m_{l+1}, k) = d_f^*(m_l, k),$$

i.e.,

$$d_{T_l}(m_l, k) = d_f^*(m_l, k).$$

In other words, the optimality of $d_T(m_{l+1}, k)$ implies the optimality of $d_T(m_l, k)$ at some time $T_l > T$, and $l-1$ becomes the largest index such that node m_{l-1} has yet to find the optimal distance to node k. By induction, there exists a time $T_0 > T$ such that

$$d_T(m_0, k) = d_f^*(m_0, k).$$

Let $t_{ik} = T_0$; then

$$d_{t_{ik}}(i, k) = d_f^*(i, k). \tag{5}$$

Lemma 2: For any $d(j, k; i)$, there exists a time $t_{jki} \geqslant T$ such that

$$d_{t_{jki}}(j, k; i) = d_f^*(j, k; i). \tag{6}$$

Proof: If $d_T(j,k;i)=d_f^*(j,k;i)$, ther. (6) is true for $t_{jki}=T$. Otherwise, the suboptimality of $d_T(j,k;i)$ and Condition 2 imply that the shortest distance from j to k via i is not direct, i.e., $k \neq i$, and

$$d_f^*(j,k;i)=s_f(j,i)+d_f^*(i,k).$$

By (5), and by Condition 5 along with the fact that current shortest distances must be transmitted whenever they change value, node j must receive distance $d_f^*(i,k)$ from node i at some time t_{jki}, yielding

$$d_{t_{jki}}(j,k;i)=d_f^*(j,k;i).$$

Lemma 3: After time T, once a current shortest distance or distance assessment becomes optimal, it remains optimal.

Proof: A current shortest distance cannot increase in the static algorithm. Condition 4 states that such a distance cannot be less than the optimum. Therefore, once optimality is attained a current shortest distance can neither increase nor decrease. A distance assessment can change only if the length of the link to the corresponding next node changes or new current shortest distance is received from that next node. Since the topology is assumed fixed after T, and current shortest distances cannot change after reaching optimality, optimal distance assessments also remain optimal.

We can now complete the proof of Theorem 1. By Lemmas 1–3, given any current shortest distance or distance assessment, there exists a time after which that particular value is guaranteed to be optimal. Let

$$T^*=\max_{i,j,k}\{t_{ij},t_{ijk}\}.$$

Then Theorem 1 is true for time T^*.

Convergence of the Dynamic Algorithm

Suppose that the dynamic algorithm discussed in Section III and detailed in Fig. 5 is used in a network, and that the assumptions preceding the description of the algorithm in that section hold for each of the lemmas, corollaries, and theorems that follow.

Lemma 4: There exists a time $\hat{t}$ such that

$$d_t(i,k;j)\geqslant d_t^*(i,k) \qquad (7)$$

for all $t\geqslant \hat{t}$, and for all i, k, and j.

Proof: Consider the topology of the network at time t_f. Since changes are not allowed after this time, the topology of the network is fixed after t_f; it consists of one or more maximal connected subnetworks, which we refer to as components. Consider an arbitrary component.

Case A: Suppose no node of the component receives or generates a local reinitialization between t' and t_f. Then the dynamic algorithm is equivalent to the static algorithm with respect to this component. The only type of topological change that can occur is link length decrease. Since (7) is assumed true for $t=t'$ and since the right-hand side of (7) can decrease but not increase, it can be proven [11] that (7) must hold for all $t\geqslant t'$.

Case B: Otherwise, at least one local reinitialization is received or generated by a node in the component. If it is generated outside the component, then the initiator has become separated from the component and a local reinitialization would be generated due to the separation, so at least one local reinitialization originated within the component between t' and t_f. Consider the final local reinitialization generated by a component node (ties can be broken arbitrarily), and say this occurs at time t_{last}. No increases, failures, or activation of new links can occur after t_{last}, the component is guaranteed to be connected after this time, and each node of the component must receive this final local reinitialization by some time t_c.

Let i be a node of the component, let $t\geqslant t_c$, and consider any $d_t(i,k;j_1)$. If $d_t(i,k;j_1)=\infty$, then clearly (7) is satisfied. Suppose, instead, that $d_t(i,k;j_1)<\infty$. If $j_1=k$, then

$$d_t(i,k;j_1)=d_t(i,k;k)=s_t(i,k)\geqslant d_t^*(i,k).$$

But if $j_1\neq k$, then

$$d_t(i,k;j_1)=s_t(i,j_1)+d_{t_1}(j_1,k),$$

where $t>t_1\geqslant t_{\text{last}}$, because for the distance from j_1 to be accepted by i, node j_1 must have received the final local reinitialization and sent it to node i. Now,

$$d_{t_1}(j_1,k)=d_{t_1}(j_1,k;j_2)$$

for some node j_2, which again is either a direct distance if $j_2=k$, or otherwise

$$d_{t_1}(j_1,k;j_2)=s_{t_1}(j_1,j_2)+d_{t_2}(j_2,k),$$

where for the same reason as before, $t_1>t_2\geqslant t_{\text{last}}$. Continuing in this manner, we see that

$$d_t(i,k;j_1)=s_t(i,j_1)+s_{t_1}(j_1,j_2)$$
$$+s_{t_2}(j_2,j_3)+\cdots+s_{t_m}(j_m,k),$$

where $t>t_1>t_2>\cdots>t_m\geqslant t_{\text{last}}$. Since only decreases in link length, and no other topological changes, can occur after t_{last}, letting $k=j_{m+1}$ we have

$$s_{t_l}(j_l,j_{l+1})\geqslant s_t(j_l,j_{l+1})$$

for all l, which implies that

$$d_t(i,k;j_1)\geqslant s_t(i,j_1)+s_t(j_1,j_2)$$
$$+s_t(j_2,j_3)+\cdots+s_t(j_m,k)$$
$$\geqslant d_t^*(i,k).$$

Thus, for each component, a time t_c can be found for which Lemma 4 is satisfied within the component. Maximizing t_c over all components yields the desired $\hat{t}$ for the entire network.

Lemma 5: No particular local reinitialization message can cause any node to reinitialize more than once.

Proof: The proof is by contradiction. Suppose that some local reinitialization, call it LRI, does cause a node to reinitialize twice. Then, of all nodes tha. reinitialize twice due to $LR1$, let node i be the first that reinitializes a second ime, and say that this occurs at time t_i. (Ties are broken arbitrarily.) This action by node i can only be caused by the receipt of $LR1$ from some neighbor, say j. Since node j s assumed not to have reinitialized twice due to $LR1$ before t_i, node j must be transmitting $LR1$ for the first time. Thus, when node i first received $LR1$, say at time t_{first}, it could not have been sent by node j.

Case A: Suppose nodes i and j were neighbors at time t_{first} and remained neighbors at least until time t_i. Receipt of $LR1$ by node i caused an embargo to be placed on node The $LR1$ message received from j by i at t_i, being the first such transmission from node j, causes the removal of the mbargo, but not the reinitialization by i that was assumed. By contradiction, nodes i and j were not neighbors at t_{first}.

Case B: If i and j were not neighbors during the entire period from t_{first} to t_i, then i and j became neighbors at ome time $t_{\text{new}} + \Delta t$, where

$$t_{\text{first}} < t_{\text{new}} + \Delta t < t_i. \qquad (8)$$

ince no node reinitializes twice before t_i, the transmission of $LR1$ by node j must be the result of a sequence of nodes, beginning with the initiator of $LR1$, each transmitting $LR1$ or the first time. This path from the initiator of $LR1$ to ode j must be loopless. If t_{LR1} is the time at which $LR1$ was generated, then by the definition of Δt,

$$t_{LR1} \leqslant t_i \leqslant t_{LR1} + \Delta t. \qquad (9)$$

Also,

$$t_{LR1} \leqslant t_{\text{first}} \leqslant t_{LR1} + \Delta t. \qquad (10)$$

Then, from (8)–(10), we get

$$t_{\text{new}} < t_i - \Delta t \leqslant t_{LR1} \leqslant t_{\text{first}} < t_{\text{new}} + \Delta t. \qquad (11)$$

In other words, $t_{\text{first}} \in [t_{\text{new}}, t_{\text{new}} + \Delta t)$, so that the index orresponding to $LR1$ is stored by node i until time $t_{\text{new}} + \Delta t$, as specified by the mechanism for adding links. Also, rom (11) we get

$$t_{\text{new}} + \Delta t < t_i < t_{\text{new}} + 2\Delta t,$$

so that at t_i, when i receives $LR1$ from j, the index orresponding to $LR1$ will be found in the set of stored ndexes, and node i will not reinitialize at t_i, again, a ontradiction. Therefore, no local reinitialization message an cause any node to reinitialize more than once.

Corollary 1: Once a particular embargo is removed from a neighbor by some node, the node can never place another mbargo with the same index on that neighbor.

Proof: An embargo can be replaced only if a second einitialization due to the same local reinitialization mes-ge occurs which, by Lemma 5, cannot happen.

Lemma 6: Every embargo that is placed is eventually emoved.

Proof: Suppose a node embargoes distance informa-n from some neighbor. The corresponding local reini-ilization must then be sent to that neighbor. If the pair of links connecting them remains active, the neighbor will receive and return the local reinitialization message, or will send this message to the node as a result of receiving the message from one of its other neighbors, thus removing the embargo. If the pair of links fail, the node removes the neighbor from its set of neighbors, eliminating the embargo. By Lemma 5, the removal of an embargo is permanent.

Lemma 7: There exists a time t_E, after which there are no embargoes anywhere in the network.

Proof: By Lemma 6, each embargo has a permanent removal time. There are a finite number of topological changes that occur between t' and t_f, thus a finite number of embargoes placed. Maximizing the removal times over all embargoes yields the desired t_E.

Lemma 8: Let $T = \max\{t_f, \hat{t}, t_E\}$. Then after time T, the dynamic algorithm is equivalent to the static algorithm.

Proof: No local reinitializations can be generated after t_f. Any embargoes that were placed have been removed by time t_E. The only operations performed after time T are the addition-comparison operations and subsequent transmission of better shortest distances, as in the static algorithm.

Theorem 2: Under the given conditions, there exists a time T^* after which each entry in the distance cube, i.e., each current shortest distance and each distance assessment, is optimal with respect to the final topology.

Proof: We individually prove that the conditions specified in Theorem 1 are satisfied.

Condition 1: $T \geqslant t_f$ implies that no topological changes occur in the network after time T.

Condition 2: By assumption, a node always knows the distance to any neighbor, and uses that distance for the distance assessment via that neighbor.

Condition 3: Consider any $d_t(i, k; j)$, where $t \geqslant T$. If node i received no local reinitializations between t' and t_f, then the only type of topological change that can occur is link length decrease. If node i constructed $d_t(i, k; j)$ based on a distance sent by node j at time $t_j < t$, then

$$\begin{aligned} d_t(i, k; j) &= s_t(i, j) + d_{t_j}(j, k) \\ &\geqslant s_t(i, j) + d_t(j, k) \end{aligned}$$

because current shortest distances cannot increase in the static algorithm. If $d_t(i, k; j)$ is not constructed from information received from j, then its value differs from $d_{t'}(i, k; j)$ only if one or more decreases in link length $s(i, j)$ occurred between t' and t. That is,

$$\begin{aligned} d_t(i, k; j) &= d_{t'}(i, k; j) - [s_{t'}(i, j) - s_t(i, j)] \\ &\geqslant s_{t'}(i, j) + d_{t'}(i, j) - s_{t'}(i, j) + s_t(i, j) \\ &\geqslant s_t(i, j) + d_t(j, k). \end{aligned}$$

If a reinitialization was received, then $d_t(i, k; j)$ is either infinite or is constructed from information sent by j, in which case, again

$$d_t(i, k; j) \geqslant s_t(i, j) + d_t(j, k).$$

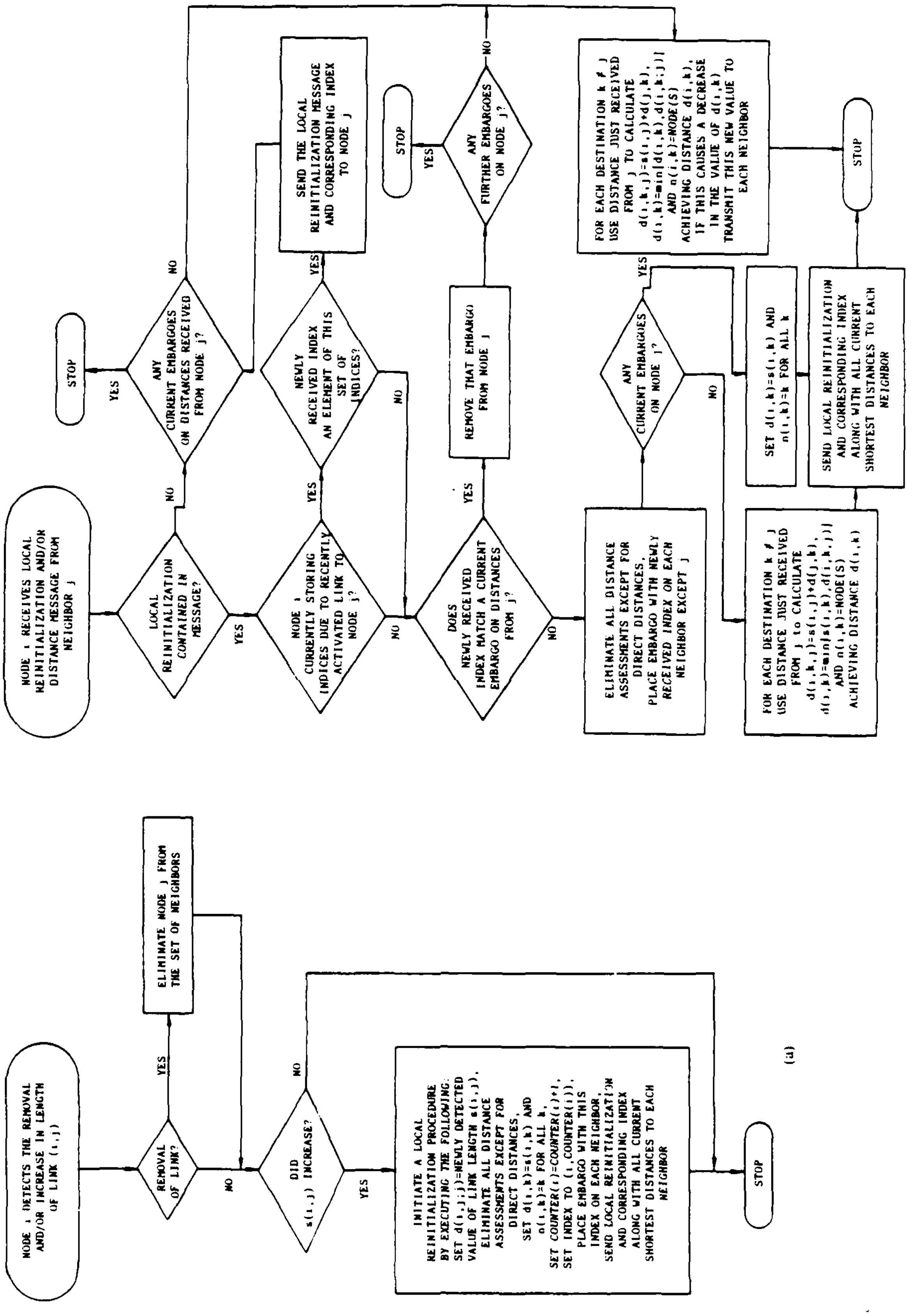
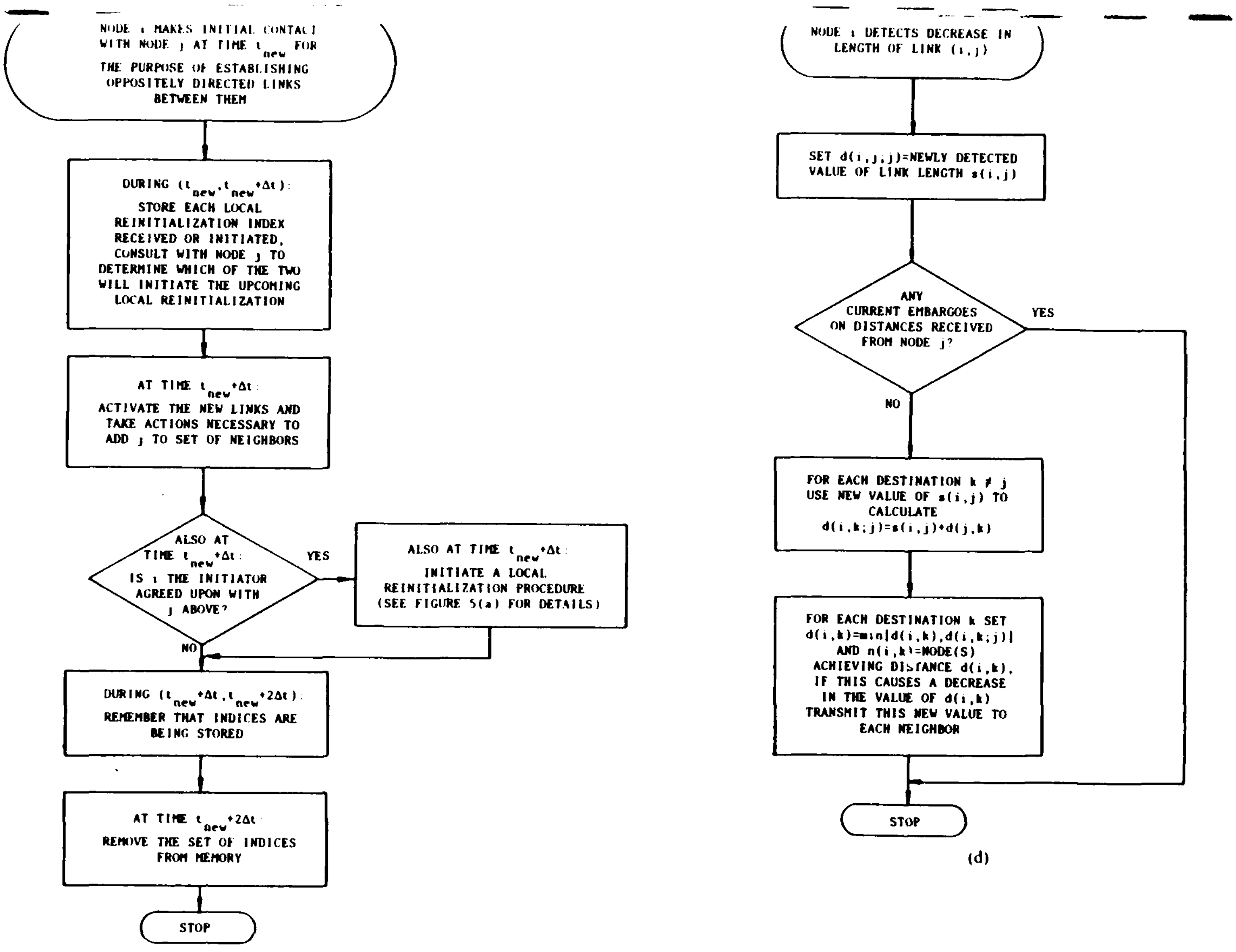

NODE i DETECTS THE REMOVAL AND/OR INCREASE IN LENGTH OF LINK (i,j)
REMOVAL OF LINK?
YES
ELIMINATE NODE j FROM THE SET OF NEIGHBORS
NO
DID s(i,j) INCREASE?
NO
YES
INITIATE A LOCAL REINITIALIZATION PROCEDURE BY EXECUTING THE FOLLOWING: SET d(i,j;j)=NEWLY DETECTED VALUE OF LINK LENGTH s(i,j). ELIMINATE ALL DISTANCE ASSESSMENTS EXCEPT FOR DIRECT DISTANCES. SET d(i,k)=s(i,k) AND n(i,k)=k FOR ALL k. SET COUNTER(i)=COUNTER(i)+1. SET INDEX TO (i,COUNTER(i)). PLACE EMBARGO WITH THIS INDEX ON EACH NEIGHBOR. SEND LOCAL REINITIALIZATION AND CORRESPONDING INDEX ALONG WITH ALL CURRENT SHORTEST DISTANCES TO EACH NEIGHBOR
STOP
(a)
NODE i RECEIVES LOCAL REINITIALIZATION AND/OR DISTANCE MESSAGE FROM NEIGHBOR j
LOCAL REINITIALIZATION CONTAINED IN MESSAGE?
NO
YES
ANY CURRENT EMBARGOES ON DISTANCES RECEIVED FROM NODE j?
YES
STOP
NO
NODE i CURRENTLY STORING INDICES DUE TO RECENTLY ACTIVATED LINK TO NODE j?
YES
NO
NEWLY RECEIVED INDEX AN ELEMENT OF THIS SET OF INDICES?
YES
NO
SEND THE LOCAL REINITIALIZATION MESSAGE AND CORRESPONDING INDEX TO NODE j
DOES NEWLY RECEIVED INDEX MATCH A CURRENT EMBARGO ON DISTANCES FROM j?
YES
NO
REMOVE THAT EMBARGO FROM NODE j
ANY FURTHER EMBARGOES ON NODE j?
YES
STOP
NO
ELIMINATE ALL DISTANCE ASSESSMENTS EXCEPT FOR DIRECT DISTANCES. PLACE EMBARGO WITH NEWLY RECEIVED INDEX ON EACH NEIGHBOR EXCEPT j
ANY CURRENT EMBARGOES ON NODE j?
YES
NO
FOR EACH DESTINATION k ≠ j USE DISTANCE JUST RECEIVED FROM j TO CALCULATE d(i,k,j)=s(i,j)+d(j,k), d(i,k)=min[d(i,k),d(i,k,j)] AND n(i,k)=NODE(S) ACHIEVING DISTANCE d(i,k). IF THIS CAUSES A DECREASE IN THE VALUE OF d(i,k) TRANSMIT THIS NEW VALUE TO EACH NEIGHBOR
STOP
SET d(i,k)=s(i,k) AND n(i,k)=k FOR ALL k
FOR EACH DESTINATION k ≠ j USE DISTANCE JUST RECEIVED FROM j to CALCULATE d(i,k,j)=s(i,j)+d(j,k), d(i,k)=min[s(i,k),d(i,k,j)] AND n(i,k)=NODE(S) ACHIEVING DISTANCE d(i,k)
SEND LOCAL REINITIALIZATION AND CORRESPONDING INDEX ALONG WITH ALL CURRENT SHORTEST DISTANCES TO EACH NEIGHBOR
STOP
(b)

Fig. 5. Actions taken by node i during dynamic algorithm for (a) generating a local reinitialization and (b) processing a message received from a neighbor


NODE i MAKES INITIAL CONTACT WITH NODE j AT TIME t_{new} FOR THE PURPOSE OF ESTABLISHING OPPOSITELY DIRECTED LINKS BETWEEN THEM
DURING (t_{new},t_{new}+Δt): STORE EACH LOCAL REINITIALIZATION INDEX RECEIVED OR INITIATED, CONSULT WITH NODE j TO DETERMINE WHICH OF THE TWO WILL INITIATE THE UPCOMING LOCAL REINITIALIZATION
AT TIME t_{new}+Δt: ACTIVATE THE NEW LINKS AND TAKE ACTIONS NECESSARY TO ADD j TO SET OF NEIGHBORS
ALSO AT TIME t_{new}+Δt: IS i THE INITIATOR AGREED UPON WITH j ABOVE?
YES
ALSO AT TIME t_{new}+Δt: INITIATE A LOCAL REINITIALIZATION PROCEDURE (SEE FIGURE 5(a) FOR DETAILS)
NO
DURING (t_{new}+Δt,t_{new}+2Δt): REMEMBER THAT INDICES ARE BEING STORED
AT TIME t_{new}+2Δt: REMOVE THE SET OF INDICES FROM MEMORY
STOP
NODE i DETECTS DECREASE IN LENGTH OF LINK (i,j)
SET d(i,j;j)=NEWLY DETECTED VALUE OF LINK LENGTH s(i,j)
ANY CURRENT EMBARGOES ON DISTANCES RECEIVED FROM NODE j?
YES
NO
FOR EACH DESTINATION k ≠ j USE NEW VALUE OF s(i,j) TO CALCULATE d(i,k;j)=s(i,j)+d(j,k)
FOR EACH DESTINATION k SET d(i,k)=min[d(i,k),d(i,k;j)] AND n(i,k)=NODE(S) ACHIEVING DISTANCE d(i,k). IF THIS CAUSES A DECREASE IN THE VALUE OF d(i,k) TRANSMIT THIS NEW VALUE TO EACH NEIGHBOR
STOP

(c)

(d)

Fig 5. (*Continued.*) (c) Activating new links and (d) detecting a branch length decrease.

Condition 4: Proof by Lemma 4.

Condition 5: Consider any $d_T(i,k)$. If this value is attained at time T, then node i must transmit this distance to each of its neighbors. If the value is attained prior to T, then $d(i,k)$ has remained constant since some time before T. Node i was required to send the distance to each of its neighbors at that time, and may have been required to make subsequent transmissions of $d_T(i,k)$ due to local reinitializations. Let t_i be the largest $t_i \leq T$ at which node i was required to transmit $d_T(i,k)$.

At t_i, node i transmits to each node that is its neighbor at that time. If any neighbor of i at time T was not a neighbor of i at t_i, then connecting links were activated between t_i and T, causing node i to reinitialize, and requiring that node i transmit $d_T(i,k)$ after the assumed final time of t_i. By contradiction, the neighbors of i at time t_i are identical to the neighbors of i in the final topology.

Suppose one of these neighbors of i, say node j, rejects the final transmission of $d_T(i,k)$ due to an embargo on node i. Since embargoes must all be removed by T, node j must receive a local reinitialization message, accompanied by distance $d_T(i,k)$, which was sent by node i after the assumed final time t_i, again a contradiction. So any node j receiving the final transmission of $d_T(i,k)$ before time T must accept the value, and have

$$\begin{aligned} d_T(j,k;i) &= s_T(j,i) + d_T(i,k) \\ &= s_f(j,i) + d_T(i,k). \end{aligned}$$

References

[1] R. Lau, R. C. M. Persiano, and P. Varaiya, "Decentralized information and control: A network flow example," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 466–473, Aug. 1972.

[2] R. G. Gallager, "A minimum delay routing algorithm using distributed computation," *IEEE Trans. Commun.*, vol. COM-25, pp. 73–85, Jan. 1977.

[3] I. W. Sandberg, "On conditions under which it is possible to synchronize digital transmission systems," *Bell Syst. Tech. J.*, vol. 48, no. 6, pp. 1999–2020, 1969.

[4] D. D. Šiljak, *Large-Scale Dynamic Systems: Stability and Structure*. New York: North-Holland, 1978.

[5] N. R. Sandell, Jr., P. Varaiya, M. Athans, and M. G. Safonov, "Survey of decentralized control methods for large scale systems," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 108–128, Apr. 1978.

[6] J. M. Abram and I. B. Rhodes, "A decentralized shortest path algorithm," in *Proc. 16th Allerton Conf. Commun., Contr. Comput.*, Univ. of Illinois, 1978, pp. 271–277.

[7] L. R. Ford, Jr., and D. R. Fulkerson, *Flows in Networks*. Princeton, NJ: Princeton Univ. Press, 1962, pp. 130–133.

[8] R. E. Kahn, "Resource-sharing computer communication networks," *Proc. IEEE*, vol. 60, pp. 1397–1407, Nov. 1972.

[9] J. M. McQuillan, G. Falk, and I. Richer, "A review of the development and performance of the ARPANET routing algorithm," *IEEE Trans. Commun.*, vol. COM-26, pp. 1802–1810, Dec. 1978.

[10] M. Schwartz and T. E. Stern, "Routing techniques used in computer communication networks," *IEEE Trans. Commun.*, vol. COM-28, pp. 539–552, Apr. 1980.

[11] J. M. Abram, "Shortest-path algorithms with decentralized information and communication requirements," D.Sc. dissertation, Dep. Syst. Sci. Math., Washington Univ., St. Louis, MO, May 1981.

[12] P. M. Merlin and A. Segall, "A failsafe distributed routing protocol," *IEEE Trans. Commun.*, vol. COM-27, pp. 1280–1287, Sept. 1979.

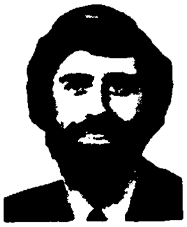
[13] D. P. Bertsekas, "Dynamic models of shortest path routing algorithms for communication networks with a ring topology," Coordinated Sci. Lab. Working Paper, Univ. of Illinois, Urbana-Champaign, Sept. 1978.

[14] J. M. McQuillan, I. Richer, and E. C. Rosen, "The new routing algorithm for the ARPANET," *IEEE Trans. Commun.*, vol. COM-28, pp. 711–719, May 1980.

Jeffrey M. Abram was born in St. Louis, MO, on August 17, 1954. He receive the B.S. degree in applied mathematics and computer science in 1975, and the M.S. and D.Sc. degrees in systems science and mathematics in 1976 and 1981, respectively, all from Washington University, St. Louis, MO.

Since June 1981 he has been a Research Engineer at Advanced Information and Decision Systems, Mountain View, CA. His current research interests include decentralized control and optimization theory.

Dr. Abram is a member of Tau Beta Pi and Pi Mu Epsilon.



Ian B. Rhodes (M'67) received the B.E. and M.Eng.Sc. degrees in electrical engineering from the University of Melbourne, Melbourne, Australia, in 1963 and 1965, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1968.

He was a faculty member at the Massachusetts Institute of Technology, Cambridge, and at Washington University, St Louis, MO, and is currently Professor of Electrical and Computer Engineering at the University of California, Santa Barbara. He has also held visiting positions at the University of Newcastle, Newcastle, Australia, and at the University of California, Berkeley. His research interests lie in the general area of system theory and its applications, with emphasis on stochastic control, communication and optimization problems, especially decentralized decision and control problems.

Dr. Rhodes has served in the past as an Associate Editor of the IEEE Transactions on Automatic Control, as Chairman of the Technical Committee on Large Systems and Differential Games of the IEEE Control Systems Society, as Associate Editor of the IFAC journal *Automatica*, and as a member of the Administrative Committee of the IEEE Control Systems Society.

Reprint of Paper:

"Smoothing Algorithms for Nonlinear Finite-Dimensional Systems", Brian D. O. Anderson and Ian B. Rhodes, *Stochastics*, **9**, pp. 139-165, 1983.

Stochastics 1983, Vol. 9, pp. 139–165
0090-9491/83/0902-0139 $18.50/0
© Gordon and Breach Science Publishers Inc., 1983
Printed in Great Britain

Smoothing Algorithms for Nonlinear Finite-Dimensional Systems

BRIAN D. O. ANDERSON†

Department of Systems Engineering, Australian National University, Canberra, ACT 2600, Australia.

and

IAN B. RHODES‡

Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

(*Accepted for publication October 26, 1982*)

Systems are considered where the state evolves either as a diffusion process or as a finite-state Markov process, and the measurement process consists either of a nonlinear function of the state with additive white noise or as a counting process with intensity dependent on the state. Fixed interval smoothing is considered, and the first main result obtained expresses a smoothing probability or a probability density symmetrically in terms of forward filtered, reverse-time filtered and unfiltered quantities; an associated result replaces the unfiltered and reverse-time filtered quantities by a likelihood function. Then stochastic differential equations are obtained for the evolution of the reverse-time filtered probability or probability density and the reverse-time likelihood function. Lastly, a partial differential equation is obtained linking smoothed and forward filtered probabilities or probability densities; in all instances considered, this equation is not driven by any measurement process. The different approaches are also linked to known techniques applicable in the linear-Gaussian case.

1. INTRODUCTION

Consider a state process x_t and an observation or measurement process z_t both evolving forward in time, with the measurement process depending in part on the state process. The task of estimating x_t from z_s, $s<t$, is termed *filtering*; the task of estimating x_t from $z_s, s<T$ for some $T>t$ is

†Work supported by the Australian Research Grants Committee. Part of this work was carried out at the address of the second author.

‡Work supported by the Office of Naval Research under Grant N00014-79-C-0549.

termed *smoothing*. Since more measurements are used in smoothing than in filtering when x_t is being estimated, smoothing leads to better estimates and is therefore to be preferred if the time delay inherent in it is acceptable.

Our main goal in this paper is to relate the smoothing problem to the filtering problem for large classes of state and measurement processes. The state processes considered are of two types: processes evolving in accordance with a diffusion equation, and continuous-time, finite-state Markov processes. For the most part, the observation processes considered are also of two types: a *nonlinear function* of the state contaminated by additive continuous-time white Gaussian noise, and a counting process, the rate of which is dependent on the state. To help understanding of our results, we also indicate briefly some corresponding discrete-time results.

Our first result (Section 3) relates the smoothing problem to the conventional filtering problem and to a *reverse-time filtering problem*: this requires the estimation of x_t given z_s for $s > t$, rather than $s < t$. Associated with this result is a second one which relates the smoothing problem to the conventional filtering problem and to a reverse-time likelihood ratio in discrete time the probability of $z_t, z_{t+1}, \ldots, z_T$ given x_t, and an analogous quantity in continuous time.

Now it is easy to note that the solution of a reverse-time filtering problem is just like that of a forward filtering problem, provided that one has a suitable reverse time signal model; *accordingly, the next* task is to explain how a reverse time signal model can be obtained from a forward-time signal model (Section 4). Such constructions happen to be available for stationary finite-state Markov processes, and for diffusion equation models, see [1 3]. We recall these results, and indicate various extensions to them to encompass the various state and observation process models of interest to us. In this way, we can obtain equations for the evolution of both forward- and reverse-filtered densities (or probabilities in the case of a finite-state process). With an equation *also available for* the corresponding unfiltered quantity, the smoothing problem is solved, at least in a formal sense.

We also present (Section 5) an equation for the evolution of the reverse-time likelihood ratio, *providing a different formal* solution to the *smoothing problem*. The solution involving the reverse-time likelihood ratio is perhaps less appealing than that involving forward- and reverse-filters, since the latter has a pleasing symmetry that reflects the symmetry of the smoothing problem itself. However, it can be that the reverse-time signal model cannot be found, and in this case one is thrown back on using the reverse time likelihood ratio

Yet a third style of equation (Section 6) relating forward filtered and smoothed probabilities or probability densities can be found. Actually, such an equation has been known for some time for *diffusion equation state* models with a measurement process containing additive white Gaussian noise, independent of the state process [4, 5]; also the equation is known for a finite-state Markov process state model with the same measurement process, see [6]. In all cases the equation is undriven by the measurement process.

We referred above to having a solution "in a formal sense" to the smoothing problem: when the state process is defined by a diffusion equation, the forward and reverse *filtering densities* are the solutions of stochastic partial differential equations, and the smoothed density is found using these two filtered densities. It is accordingly of interest to indicate situations when a finite-dimensional solution to the smoothing problem can be obtained. Besides the standard linear-Gaussian problem, and the problem with a finite-state Markov process state model, we describe one such situation in Section 6 involving a linear diffusion equation state model and an observation model containing a mixture of linear-Gaussian observations and space-time counting process *observations*.

Besides the references noted above on nonlinear smoothing, there have been a number of other contributions (see, e.g. [7 12]) including some which are very significant. None, as far as we are aware, has attempted to use forward and reverse-time filters in a symmetric way, nor has developed, for as comprehensive a set of state and observation models as considered here, the equation for the evolution of the reverse-time likelihood ratio and the equation relating the smoothed and forward filtered densities which is undriven by the measurements.

The general thrust of this paper is to suggest a wide variety of algorithms, rather than to present the fine details governing an existence proof associated with one algorithm. To the extent that the different competing algorithms for linear-Gaussian smoothing are now all apparently captured in a non-linear framework, we may therefore also be setting out a representative collection of bases upon which to build finite-dimensional (near- but non-optimal) smoothers.

2. NOTATION AND BACKGROUND ASSUMPTIONS

Throughout the paper we shall consider stochastic processes defined on a fixed measurable space (Ω, F). Unless otherwise indicated, all processes are considered to be *defined on the fixed interval* $[0, T]$ *and to be* F_t-adapted, where $\{F_t, t \in [0, T]\}$ is an increasing family of sub-σ-algebras of F. Mostly, we shall be concerned with continuous-time processes in which $[0, T]$ has its usual meaning; however, we shall also consider briefly discrete-time processes in which case $t, T \in Z_+$, the set of positive integers. We denote by $X \triangleq \{x_t, F_t, P\}$ a finite-dimensional *state* process which is F_t-adapted, and P-measurable on (Ω, F) under the probability measure P. We denote by $\sigma(x_t)$ the sub-σ-algebra of F_t generated by the random variable x_t, and by X_t and X_{t+} the sub-σ-algebras of F_t generated by the collections of random variables $\{x_s; s \in [0, t)\}$ and $\{x_s; s \in (t, T]\}$.

We shall concentrate on two x_t-process models a diffusion model D, and a finite state Markov model F, defined as follows: First,

$$D \qquad dx_t = f(x_t, t)\,dt + g(x_t, t)\,dv_t \tag{2.1}$$

where $V = \{v_t, F_t, P\}$ is a Wiener process taking values in R^p, f and g satisfy appropriate smoothness and rate-of-growth conditions, and x_0 is independent of $\{v_t\}$.

The finite-state model F is defined as being a Markov process such that if p_t is an n-vector with i-th entry the probability that x_t is at the i-th of n levels, then

$$F \qquad dp_t = A(t)p_t\,dt \tag{2.2}$$

The matrix $A(t)$ satisfies

$$a_{ij}(t) \geqq 0 \quad \text{for} \quad i \neq j, \sum_{k=1}^{n} a_{ki}(t) = 0. \tag{2.3}$$

It is, of course, possible to consider combinations of D and F, and the subsequent ideas and results of the paper can be extended to such combined models.

Associated with the state process X there will also be an *observation or measurement* process $Z = \{z_t, F_t, P\}$, which is linked to x_t by some form of observation model. Like the x-process, Z is finite-dimensional, F_t-adapted, and P-integrable on (Ω, F) under the probability measure P. In slight contrast to our definition of X_t and X_{t+} we define Z_t and Z_{t+} to be the sub-σ-algebras of F_t generated by the increments of z_s in the intervals $[0, t)$ and $[t, T)$. We shall be interested in a model G with additive white Gaussian measurement noise and a model C with z a counting process whose intensity depends on the state. In more detail

$$G \qquad dz_t = h(x_t, t)\,dt + M(t)\,dw_t, \qquad z_0(\omega) = 0, \tag{2.4}$$

where $W = \{w_t, F_t, P\}$ is a Wiener process taking values in R^m, $M(t): R^m \to R^m$ is deterministic and invertible, and h is jointly measurable in x_t and t. The process W is independent of X and V and

$$R(t) \triangleq M(t)M'(t) \tag{2.5}$$

is uniformly bounded with a uniformly bounded inverse, and L_2-integrable on $[0, T]$. Further

$$\int_0^t \|h(x_t, t)\|^2_{R^{-1}(t)}\,dt < \infty \qquad P\text{-a.s.} \tag{2.6}$$

Next, the counting process model is

$$C \qquad dz_t = dN_t \quad z_0(\omega) = 0, \tag{2.7}$$

where $N = \{N_t, F_t, P\}$ is a counting process with intensity $\Lambda = \{\lambda_t, F_t, P\}$ satisfying

$$\lambda_t = h(x_t, t), \tag{2.8}$$

where h is jointly measurable in x_t and t, $\int_0^t h(x_s, s)\,ds < \infty$ P-a.s., and h is such that Λ is predictable and positive a.s. Several calculations involving the appropriate Ito rule also require λ_t to be uniformly bounded a.s.; for convenience we take this bound to be unity, since the extension to arbitrary upper bound is straightforward [see, e.g., 11].

A crucial and well-known consequence of the above assumptions is:

2.1. Observation

With either state model D or F and either observation model G or C, the σ-algebra $X_t \vee Z_t$ is conditionally independent given x_t of $X_{t+} \vee Z_{t+}$. Here $X \vee Z$ denotes the least σ-algebra containing X and Z.

In the sequel, we shall be interested in the probabilities (model F) or probability densities (model D) $p(x_t)$, $p(x_t|Z_t)$, $p(x_t|Z_{t+})$, and $p(x_t|Z_T)$. We shall refer to these as unfiltered, filtered or forward filtered, reverse filtered, and smoothed probabilities or probability densities, respectively. Our concern is not to pay close attention to conditions for the existence of these densities, which by and large have been dealt with exhaustively in other works; for the most part, we shall simply assume that the conditions are fulfilled which ensure their existence. At times, however, it would be satisfactory to work with distributions.

3. BASIC SMOOTHING FORMULAS

The various smoothing formulas we first obtain can be thought of as taking one of two forms:

$$p(x_t | Z_T) = \frac{p(x_t | Z_{t-})p(x_t | Z_{t+})}{N p(x_t)} \tag{3.1}$$

and

$$p(x_t | Z_T) = \frac{p(x_t | Z_{t-})}{N'} \cdot [\text{Object like } p(Z_{t+} | x_t)]. \tag{3.2}$$

Here, N and N' are normalizing quantities, and so are Z_T-dependent, but not x_t-dependent.

We can obtain a feel for these formulas by considering a discrete-time version of one of the models given earlier. Thus, suppose x_t is a continuous-state, discrete-time, Markov-process, and $z_t = h(x_t) + n_t$, where n_t is discrete-time white noise, with the $\{n_t\}$ and $\{x_t\}$ processes independent. Let Z_{t-} denote $\{z_0, z_1, \ldots, z_t\}$, Z_{t+} denote $\{z_{t+1}, \ldots, z_T\}$, and Z_T denote $\{z_0, z_1, \ldots, z_T\}$. Then Observation 2.1 becomes simply

$$p(Z_T | x_t) = p(Z_{t-} | x_t)p(Z_{t+} | x_t). \tag{3.3}$$

Now, assuming all densities exist, we have from Bayes' rule and (3.3)

$$p(x_t | Z_T) = \frac{p(Z_T | x_t)p(x_t)}{p(Z_T)} = \frac{p(Z_{t-} | x_t)p(Z_{t+} | x_t)p(x_t)}{p(Z_T)}$$

$$= \frac{p(Z_{t-} | x_t)p(Z_{t+} | x_t)p(x_t)p(x_t)}{p(Z_{t-})p(Z_{t+})p(x_t)} \cdot \frac{p(Z_{t-})p(Z_{t+})}{p(Z_T)}$$

and (3.1) is immediate upon reapplication of Bayes' rule and identification of $p(Z_T)[p(Z_{t-})p(Z_{t+})]^{-1}$ as the normalizing quantity N. To derive (3.2), the last equality in the algebra above is replaced by

$$p(x_t | Z_T) = \frac{p(Z_{t-} | x_t)p(x_t)}{p(Z_{t-})} \cdot \frac{p(Z_{t-})}{p(Z_T)} \cdot p(Z_{t+} | x_t),$$

which is (3.2) with $N' = p(Z_T)p^{-1}(Z_{t-})$ and the "object like $p(Z_{t+} | x_t)$" simply $p(Z_{t+} | x_t)$ itself. If x_t is a discrete-time finite-state process, the same calculations yield (3.1) and (3.2) where the p symbols should be interpreted as probabilities rather than probability densities.

The above Bayes' rule calculations cannot be duplicated exactly for continuous-time processes. Nevertheless, tools for capturing the Bayes' idea exist in the form of representation theorems (see, e.g. [13, 14]). For the observation process G, define

$$L_{s_1, s_2} = \exp\left\{\int_{s_1}^{s_2} h'(x_s, s)R^{-1}(s)\,dz_s - \tfrac{1}{2}\int_{s_1}^{s_2} \|h(x_s, s)\|^2_{R^{-1}(s)}\,ds\right\} \tag{3.4}$$

and for observation process C, define

$$L_{s_1, s_2} = \exp\left\{\int_{s_1}^{s_2} \ln h(x_s, s)\,dz_s - \int_{s_1}^{s_2} [h(x_s, s) - 1]\,ds\right\} \tag{3.5}$$

Then we have for any combination of state and observation models the representation theorem (see, e.g., [11], [15])

$$p(x_s = x | Z_{t-}) = \frac{E_X[L_{0,t} | Z_{t-}, x_s = x]}{E_X[L_{0,t} | Z_{t-}]} p(x_s = x), \tag{3.6}$$

for all $s, t \in [0, T]$. The notation $E_X[L_{0,t} | Z_{t-}]$ or $E_X[L_{0,t} | Z_{t-}, x]$ means that the expectation is to be taken with respect to the process X with Z_t fixed, i.e. the equality in (3.6) holds pointwise for each observation trajectory (Z_{t-}). We note that (3.6) is in the form of the Bucy representation theorem [15] involving only one probability measure, in contrast to representation theorems in which the two conditional expectations in a form similar to (3.6) are with respect to a transformed probability measure.† We note, too, that one way to derive (3.6) is via measure transformations [see 11].

One can think of $E_X[L_{0,t} | Z_{t-}]$ and $E_X[L_{0,t} | Z_{t-}, x_s = x]$ as being like $p(Z_{t-})$ and $p(Z_{t-} | x_s)$, and then (3.6) becomes Bayes' theorem. This thinking, and the discrete time derivations, guide the following development.

Observe with the aid of (3.4) and (3.5) that

$$L_{0,T} = L_{0,t}L_{t,T}. \tag{3.7}$$

†These versions involving transformed measures hold under conditions that are weaker than our standing assumptions. Pardoux [7, 8, 16] has used a representation involving transformed measures to derive an expression associated with the "asymmetric form" (3.2) of the smoothed density that, for state model D and observation model G, is more general than Eq. (5.9) in Section 5 in that it includes correlated state and observation noise processes v and w.

Accordingly, using (3.4) and (3.5) again along with Observation 2.1,

$$E_X[L_{0,T}|Z_{T}, x_t = x] = E_X[L_{0,t}|Z_{t}, x_t = x]\, E_X[L_{t,T}|Z_{t,T}, x_t = x]. \quad (3.8)$$

With s and t in (3.6) identified with t and T, we then have

$$p(x_t = x|Z_{T}) = \frac{E_X[L_{0,T}|Z_T, x_t = x)}{E_X[L_{0,T}|Z_T]}\, p(x_t = x)$$

$$= \frac{E_X[L_{0,t}|Z_{t}, x_t = x]\, p(x_t = x)}{E_X[L_{0,t}|Z_{t}]}$$

$$\frac{E_X[L_{t,T}|Z_{t,T}, x_t = x]\, E_X[L_{0,t}|Z_{t}]}{E_X[L_{0,T}|Z_T]}$$

$$= p(x_t = x|Z_{t})\, E_X[L_{t,T}|Z_{t,T}, x_t = x]\cdot \frac{E_X[L_{0,t}|Z_{t}]}{E_X[L_{0,T}|Z_T]} \quad (3.9)$$

The second equality follows by using (3.7) and the third one using (3.6) with s and t in (3.6) both equal to t. Eq. (3.9) is the concrete form of (3.2), where of course $E_X[L_{t,T}|Z_{t,T}, x_t = x]$ is the "object like $p(Z_{t,T}|x_t)$" and the inverse of the third factor on the right side is the normalizing quantity N'. An equation like this (but involving expectations under a transformed probability measure) is studied by Pardoux in [16] for state model D and observations G and C where a wide range of precise assumptions are set out under which the equation is valid.

To obtain (3.1), notice that if the origin is shifted to time t and we then identify s with t and t with T in (3.6) we have

$$p(x_t = x|Z_{t,T}) = \frac{E_X[L_{t,T}|Z_{t,T}, x_t = x]}{E_X[L_{t,T}|Z_{t,T}]}\, p(x_t = x),$$

and when this is used to substitute for the second term on the right side of (3.9), we obtain

$$p(x_t = x|Z_{T}) = \frac{p(x_t = x|Z_{t})\, p(x_t = x|Z_{t,T})}{p(x_t = x)}$$

$$\frac{E_X[L_{0,t}|Z_{t}]\, E_X[L_{t,T}|Z_{t,T}]}{E_X[L_{0,T}|Z_T]} \quad (3.10)$$

This is (3.1), with a more precise identification of the normalizing quantity N.

The above arguments apply for either observation model G or C, and either state model D or F, save that p must be interpreted for model F as a probability and for model D as a probability density. We note that (3.1) and a version of (3.2) have been derived via an alternative route in [21, Section 6.7] for state model D and observation model C.

4. REVERSE FILTERED DENSITIES

A key conclusion of the preceding section was the formula (3.1) expressing the smoothed probability or probability density symmetrically in terms of a forward filtered, a reverse filtered, and an unfiltered probability or probability density. In this section, we note how equations for the reverse filtered probability or density can be derived.

Let us focus for the moment on the state model D, repeated as

$$dx_t = f(x_t, t)\,dt + g(x_t, t)\,dv_t. \quad (4.1)$$

The properties of this model include independence of x_0 and v_t, and, more generally, independence of x_t and $v_s - v_t$ for any $s > t$, *but not for* $s < t$, i.e. present and past states are independent of future noise, but present and future states are not independent of past noise. The model is thought of as evolving forward in time, and (4.1) is understood as defining a forward Ito equation. What is needed for the calculation of the reverse filtered density is a signal model that evolves backwards in time with the property that, relative to this reverse time evolution, "future" noise is independent of "past" signal.

The main result of [3] is that if $p(x_t, t)$ exists for all t and x, then one may define

$$(dv_t^r)^k = (dv_t)^k + \frac{1}{p(x,t)}\cdot\left\{\sum_j \frac{\partial}{\partial x_t^j}[p(x_t,t)g^{jk}(x_t,t)]\,dt\right\} \quad (4.2)$$

and

$$[f^r(x_t,t)]^i = [f(x_t,t)]^i \quad \frac{1}{p(x_t,t)}\left\{\sum_{j,k} \frac{\partial}{\partial x_t^j}[p(x_t,t)g^{ik}(x_t,t)g^{jk}(x_t,t)]\right\} \quad (4.3)$$

with the following properties:

$$v_t^r \text{ is a Wiener process} \quad (4.4)$$

$$dx_t \overset{b}{=} f^r(x_t,t)\,dt + g(x_t,t)\,dv_t^r, \quad (4.5)$$

where (4.5) is a backward Ito equation, i.e., the integral form of (4.5) involves a backward Ito integral, and

$$x_t \quad \text{and} \quad v_s^r \quad v_t^r \text{ are independent for } s < t. \quad (4.6)$$

This means that if t is the present, and $s>t$ is called the "past" for the backwards-evolving (4.5) and $s<t$ the "future", then "future" driving noise is independent of present and past state. (Henceforth, b or f above the equality sign will stress that the equation is a backward or forward equation.)

The above construction is an extension of one in [17] for the linear-Gaussian case.

Just as the state model can be conceived as evolving in reverse time, so can a measurement model. Under our assumptions, the noise W in the measurement model G of (2.4) is independent of the state process X for either state model D or F. In this case the appropriate backward measurement model is just (2.4) itself, i.e.

$$dz_t = h(x_t, t)\,dt + M(t)\,dw_t, \tag{4.7}$$

though (4.7) has two interpretations, as a forward and as a backward equation, i.e.

$$z_{t+dt} - z_t \overset{f}{=} h(x_t, t)\,dt + M(t)[w_{t+dt} - w_t] \qquad (4.7f)$$

and

$$z_t - z_{t\;\;dt} \overset{b}{=} h(x_t, t)\,dt + M(t)[w_t - w_{t\;\;dt}]. \qquad (4.7b)$$

Had the driving noise V in model D been correlated with the observation noise W in model G, the backward observation model would have differed from the forward model to reflect appropriate correlation with the process $\{v_t^r\}$ in (4.5).

The point of these observations is that they allow the setting up of equations for the reverse filtered density that are just like those for the forward filtered density. For example, for the state model defined by (4.1) and the measurement Eq. (2.4), the unnormalized forward filtered density $p_{fu}(x_t)$ is given from the Zakai [18] equation as, (see, e.g., [7, 8]),

$$dp_{fu} \overset{f}{=} \mathscr{L}[p_{fu}]\,dt + p_{fu}h'R^{\;1}\,dz_t, \tag{4.8}$$

where

$$\mathscr{L}(\cdot) = \;\sum_i \frac{\partial}{\partial x_t^i}[f^i(x_t, t)\cdot] + \tfrac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_t^i\,\partial x_t^j}\{p[g(x_t, t)g'(x_t, t)]^{ij}\cdot\} \tag{4.9}$$

and $p_{fu}(x_0 \mid Z_0\;\;) = p(x_0)$. Accordingly, we see from (4.5) and (4.7b) that the equation for the associated unnormalized reverse-time filtered density is simply

$$dp_{ru} \overset{b}{=} \;\;\mathscr{L}_r[p_{ru}]\,dt \;\; p_{ru}h'R^{\;1}\,dz_t, \tag{4.10}$$

with $\mathscr{L}_r$ defined like $\mathscr{L}$, save that $[\;-f^r]^i$ replaces $[f]^i$ in the definition (4.9); the boundary condition is, naturally, $p_{ru}(x_T \mid Z_{T+}) = p(x_T)$.

The general approach to obtain (4.10) is simple. One has backward models for the state process and measurement process, one makes a change of time variable so that forward models of related processes are obtained, one uses (4.8) to write down the forward unnormalized filtered density, and then one changes the time variable again, to get (4.10). The full details are set out in Appendix A, which also serves as a guide to similar subsequent calculations that are omitted.

As pointed out by a reviewer, this equation can almost certainly be derived from results of [16, Eq. (3.15)], since $p(x_t = x \mid Z_{t+})$ is given in terms of $E_X[L_{T,t} \mid Z_{t+}, x_t = x]$ just above (3.10). However, such a derivation obscures the essential time symmetry of the smoothing problem, which underpins the development of this section.

Instead of working with the unnormalized equation, one can work if desired with the normalized equation for $p_f = p(x_t \mid Z_t\;\;)$. A version for scalar x_t and z_t may be found in [6]. Alternatively, it may be obtained in a similar manner to a calculation of [19]. The answer is standard, but we set out the brief calculation as a guide to later constructions. Let

$$n_f(t) = \int_{R^n} p_{fu}(x_t, t)\,dx_t, \tag{4.11}$$

where x_t is an n-vector. Then (4.8) implies

$$dn_f(t) \overset{f}{=} \hat{h}_f'\,R^{\;1}\,dz_t n_f(t), \tag{4.12}$$

where $\hat{h}_{ft} = E[h(x_t, t) \mid Z_t\;\;]$, and then the Ito rules give

$$d\left(\frac{p_{fu}(x_t, t)}{n_f(t)}\right) = \frac{dp_{fu}}{n_f}\;\;\left(\frac{p_{fu}}{n_f}\right)\left(\frac{dn_f}{n_f}\right)\;\;\left(\frac{dp_{fu}}{n_f}\right)\left(\frac{dn_f}{n_f}\right) + \frac{p_{fu}}{n_f}\left(\frac{dn_f}{n_f}\right)\left(\frac{dn_f}{n_f}\right) \tag{4.13}$$

which leads to

$$dp_f \overset{f}{=} \mathscr{L}(p_f)\,dt + (h - \hat{h}_f)'R^{\;1}(dz_t \;\; \hat{h}_f\,dt). \tag{4.14}$$

The corresponding reverse equation is, with $\hat{h}_r = E[h \mid Z_{t+}]$,

$$dp_r \overset{b}{=} \;\;\mathscr{L}_r(p_r)\,dt \;\; (h \;\; \hat{h}_r)'R^{\;1}(dz_t \;\; \hat{h}_r\,dt). \tag{4.15}$$

Now let us turn to the finite-state Markov model F. The construction of a reverse model given a stationary forward model is undertaken in [1, 2]. Here, we note an extension. Suppose that p_t, $t\in[0,T]$, is the solution of (2.2). Suppose further that the entries of p_t are all identically nonzero for t in $[0,T]$. (It is easy to see that this is equivalent to demanding that all entries of p_0 are nonzero.) Define

$$\Pi_t = \operatorname{diag}[p_{1t},\ldots,p_{nt}], \tag{4.16}$$

there being n different state levels. Also, define

$$A^r(t) = \Pi_t A'(t)\Pi_t^{-1} - \dot{\Pi}_t\Pi_t^{-1}. \tag{4.17}$$

Then the reverse model is a finite state Markov process $x^r(\cdot)$ and, with the i-th entry of p^r_t denoting the probability that $x^r(t)$ is in level i,

$$dp^r_t = -A^r(t)p^r_t\,dt \qquad p^r_T = p_T. \tag{4.18}$$

It is quickly verified that for all t the entries of A^r satisfy $a^r_{ij}\geqq 0$ if $i\neq j$, and that $\sum_{i=1}^n a^r_{ij}=0$ for all j, as required for (4.18) to be a (backward) finite-state process. To see that (4.18) is, in fact, the reverse model associated with (2.2), it is necessary and sufficient to show that, for all i, j and t,

$$Pr[x(t)=i] = Pr[x^r(t)=i] \tag{4.19}$$

and

$$Pr[x(t)=i\,|\,x(s)=j] = Pr[x^r(t)=i\,|\,x^r(s)=j]. \tag{4.20}$$

Now, it is not hard to verify that if $\Phi(t,s)$ is the transition matrix which satisfies

$$\frac{d}{dt}\Phi(t,s) = A(t)\Phi(t,s) \qquad \Phi(s,s)=I \tag{4.21}$$

for all t and s, then $\Psi(t,s)=\Pi_t\Phi'(s,t)\Pi_s^{-1}$ satisfies

$$\frac{d}{dt}\Psi(t,s) = -A^r(t)\Psi(t,s) \qquad \Psi(s,s)=I, \tag{4.22}$$

for all t and s. Also, observe that $[1,1,\ldots,1]\Phi(t,s)=[1,1,\ldots,1]$ for all t and s; for if p_s is the probability vector at time s, that at time t is $\Phi(t,s)p_s$ and the entries of this vector must sum to 1 for all probability vectors p_s.

Now, to establish (4.19), observe that

$$p^r_t = \Psi(t,T)p^r_T = \Pi_t\Phi'(T,t)\Pi_T^{-1}p_T = \Pi_t\Phi'(T,t)[1,1,\ldots,1]' = \Pi_t[1,1,\ldots,1]' = p_t.$$

Now consider (4.20). Let e_j denote a unit vector with 1 in the j-th position. We have, if $t>s$,

$$\begin{aligned} p[x^r_t=i\,|\,x^r_s=j] &= \frac{p[x^r_s=j\,|\,x^r_t=i]\,p[x^r_t=i]}{p[x^r_s=j]} = \frac{e'_j\Psi(s,t)e_ip_{it}}{p_{js}} \\ &= e'_j\Pi_s^{-1}\Psi(s,t)\Pi_te_i = e'_j\phi'(t,s)e_i \\ &= e'_i\phi(t,s)e_j = p[x_t=i\,|\,x_s=j]. \end{aligned}$$

A similar argument works for $t<s$ and thus establishes (4.20).

Suppose that when x_t is in the i-th state, $h(x_t,t)=h_i(t)$. Suppose also that the observation model G applies, and let p_{if} denote the i-th entry of the forward filtered probability vector. Thus the forward filter equations are (expanding [20] to the time-varying, multiple-output, case and omitting the common time argument t)

$$dp_{if} \overset{f}{=} \sum_j a_{ij}p_{jf}\,dt + [h_i - \sum_{j=1}^n h_jp_{jf}]'R^{-1}[dz_t - \sum_{j=1}^n h_jp_{jf}\,dt]p_{if}. \tag{4.23}$$

The observation model is the same in reverse time, and so in view of the equality of all joint densities associated with the forward and reverse models, we see also that

$$dp_{ir} \overset{b}{=} -\sum_j a^r_{ij}p_{jr}\,dt - [h_i - \sum_{j=1}^n h_jp_{jr}]'R^{-1}[dz_t - \sum_{j=1}^n h_jp_{jr}\,dt]p_{ir}. \tag{4.24}$$

We also note the unnormalized form of these equations:

$$dp_{ifu}(t) \overset{f}{=} \sum_j a_{ij}(t)p_{jfu}(t)\,dt + h_i(t)R^{-1}(t)\,dz_tp_{ifu}(t) \tag{4.25}$$

$$dp_{iru}(t) \overset{b}{=} -\sum_j a^r_{ij}(t)p_{jru}(t)\,dt - h_i(t)R^{-1}(t)\,dz_tp_{iru}(t). \tag{4.26}$$

We note that the connection between (4.25) and (4.23) follows easily on using the normalizing factor

$$n_f(t) = \sum_i p_{ifu}(t),$$

which satisfies

$$dn_j(t) \overset{f}{=} \left[\sum_{i=1}^{n} h_i(t) \frac{p_{ijfu}(t)}{n_j(t)} \right]' R^{-1}(t)\, dz_t n_j(t). \tag{4.27}$$

The use of (4.13) yields (4.23) with $p_{if} = p_{ifu}/n_j$.

Now consider observation model C. When the state model is D, the forward normalized filter equation is [11]

$$dp_f \overset{f}{=} \mathscr{L}[p_f]\,dt + [h - \hat{h}_f][\hat{h}_f]^{-1}[dz_t - \hat{h}_f\, dt]p_f. \tag{4.28}$$

The observation model is evidently the same in reverse time as in forward time save for its interpretation as a backwards rather than a forwards equation. Consequently, using arguments analogous to those in Appendix A, the reverse equation for $p_r = p(x_t | Z_{t,T})$ is

$$dp_r \overset{b}{=} -\mathscr{L}_r[p_f]\,dt - [h - \hat{h}_r][\hat{h}_r]^{-1}[dz_t - \hat{h}_r\, dt]p_r. \tag{4.29}$$

We assert that the unnormalized equation corresponding to (4.28) is

$$dp_{fu} \overset{f}{=} \mathscr{L}(p_{fu})\,dt + (h - 1)(dz_t - dt)p_{fu} \tag{4.30}$$

which implies that

$$dp_{ru} \overset{b}{=} -\mathscr{L}_r(p_{ru})\,dt - (h - 1)(dz_t - dt)p_{ru}. \tag{4.31}$$

The verification that (4.30) leads to (4.28) proceeds in a parallel fashion to the derivation of (4.14) from (4.8) via (4.11) (4.13) using, in this case, the Ito differential rule for counting process observations that can be found, for example, in [11], [21].

Finally, for the combination of observation model C and state model F, we have for the normalized Eq. [11]

$$dp_{if} \overset{f}{=} \sum_j a_{ij}p_{jf}\, dt + [h_i(t) - \hat{h}_f(t)][\hat{h}_f(t)]^{-1}[dz_t - \hat{h}_f(t)\, dt]p_{if} \tag{4.32}$$

and for the unnormalized equation

$$dp_{ifu} \overset{f}{=} \sum_j a_{ij}p_{jfu}\, dt + (h - 1)(dz_t - dt)p_{ifu}. \tag{4.33}$$

The reverse equations are

$$dp_{ir} \overset{b}{=} -\sum_j a'_{ij}p_{jr}\, dt - [h_i - \hat{h}_r][\hat{h}_r]^{-1}[dz - \hat{h}_r\, dt]p_{ir} \tag{4.34}$$

and

$$dp_{iru} \overset{b}{=} -\sum_j a'_{ij}p_{jru}\, dt - (h_i - 1)(dz_t - dt)p_{iru}. \tag{4.35}$$

The derivation of the last three equations also proceeds in a similar fashion to the earlier derivations.

5. REVERSE LIKELIHOOD FUNCTION

Recall that one of the two basic smoothing formulas derived in Section 3 involved $p(Z_{t,T} | x_t)$ for discrete time and $E_X[L_{t,T} | Z_{t,T}, x_t = x]$ for continuous time. In this section, we shall present differential equations for the evolution of this latter quantity, which is in effect a likelihood function.

For state models of type D and observation models of type G, an equation for this quantity has been derived in [7,8]. We begin with a different derivation for this case, the form of which permits easy extension to state models F and observation models C. Define

$$\phi_t = \int_t^T h'(x_s, s)R^{-1}(s)\, dz_s - \tfrac{1}{2}\int_t^T \|h(x_s, s)\|^2_{R^{-1}(s)}\, ds \tag{5.1}$$

so that

$$d\phi_t \overset{f}{=} -h'(x_t, t)R^{-1}(t)\, dz_t + \tfrac{1}{2}\|h(x_t, t)\|^2_{R^{-1}(t)}\, dt, \tag{5.2}$$

where this is a *forward* Ito equation. By the standard Ito rule applied to the combined process $[\phi_t, x'_t]'$, it is straightforward to derive

$$dL_{t,T} = d(\exp\phi_t) \overset{f}{=} -L_{t,T}h'(x_t, t)R^{-1}(t)\, dz_t + L_{t,T}\|h(x_t, t)\|^2_{R^{-1}(t)}\, dt. \tag{5.3}$$

To obtain a backward equation, one uses double the adjustment required to form the Stratonovich equation; the adjustment can be obtained by applying it to the Markov process $[L_{t,T}, x'_t]'$. We obtain

$$dL_{t,T} \overset{b}{=} -L_{t,T}h'(x_t, t)R^{-1}\, dz_t$$

as a backward equation, so that

$$L_{t,T} \overset{b}{=} \int_t^T L_{s,T}\, h'(x_s, s)R^{-1}(s)\, dz_s + 1. \tag{5.4}$$

Then

$$E[L_{t,T} \mid Z_{t}, x_t = x] \overset{b}{=} \int_t^T \left\{ \int_{R^n} E[L_{s,T} \mid Z_{s}, x_s] h'(x_s, s) \times R^{-1}(s) p(x_s \mid x_t = x)\, dx_s \right\} dz_s + 1. \qquad (5.5)$$

Let us define

$$V_t(x) = E[L_{t,T} \mid Z_{t}, x_t = x]. \qquad (5.6)$$

Then (5.5) yields

$$dV_t(x_t) \overset{b}{=} -V_t(x_t) h'(x_t, t) R^{-1}(t)\, dz_t + \int_t^T \left\{ \int_{R^n} V_s(x_s) h'(x_s, s) R^{-1}(s) \frac{\partial}{\partial t} p(x_s \mid x_t = x)\, dx_s \right\} dz_s. \qquad (5.7)$$

Recall the standard backward Kolmogorov Eq., [13]:

$$\frac{\partial}{\partial t} p(x_s \mid x_t) = -\mathcal{L}^a[p(x_s \mid x_t)] \quad (s \geqq t), \qquad (5.8)$$

where $\mathcal{L}^a$ is the adjoint of $\mathcal{L}$ defined in (4.10), and the x-differentiation in $\mathcal{L}^a$ is with respect to x_t, not x_s. Then, using (5.8) in (5.7) yields

$$dV_t(x_t) \overset{b}{=} -V_t(x_t) h'(x_t, t) R^{-1}(t)\, dz_t$$
$$-\mathcal{L}^a \int_t^T \left\{ \int_{R^n} V_s(x_s) h'(x_s, s) R^{-1}(s) p(x_s \mid x_t = x)\, dx_s \right\} dz_s.$$

The last term can be replaced, via (5.5), with $-\mathcal{L}^a[V_t(x) - 1]\, dt = -\mathcal{L}^a[V_t(x)]\, dt$. Thus

$$dV_t(x_t) \overset{b}{=} -\mathcal{L}^a[V_t(x_t)]\, dt - V_t(x_t) h'(x_t, t) R^{-1}(t)\, dz_t. \qquad (5.9)$$

Eq. (5.9), with the boundary condition $V_T(x_T) = 1$, is the equation sought. As mentioned in an earlier footnote, this equation has been derived by Pardoux [16, Eq. (3.15)].

Suppose now that the state model is the finite-state model F, and the observation model is type G. Then, a similar argument to that leading to (5.5) gives

$$E[L_{t,T} \mid Z_{t}, x_t = i] \overset{b}{=} \int_t^T \sum_j E[L_{s,T} \mid Z_{s}, x_s = j] \times h'_j(s) R^{-1}(s) p(x_s = j \mid x_t = i)\, dz_s + 1. \qquad (5.10)$$

Define, in similar vein to (5.6),

$$V_i(t) = E[L_{t,T} \mid Z_{t}, x_t = i].$$

Observe that, with $\Phi(t, s)$ the transition matrix associated with $\dot{x} = A(t)x$,

$$p(x_s = j \mid x_t = i) = e'_j \Phi(s, t) e_i$$

and

$$\frac{\partial}{\partial t} p(x_s = j \mid x_t = i) = \frac{\partial}{\partial t} e'_j \Phi^{-1}(t, s) e_i = -e'_j \Phi^{-1}(t, s) A(t) e_i$$
$$= -e'_j \Phi(s, t) \sum_{k=1}^n e_k e'_k A(t) e_i = -\sum_{k=1}^n p(x_s = j \mid x_t = k) a_{ki}(t).$$

Consequently, differentiation of (5.10) yields

$$dV_i(t) \overset{b}{=} \int_t^T \sum_j V_j(s) h'_j(s) R^{-1}(s) \left[-\sum_{k=1}^n p(x_s = j \mid x_t = k) a_{ki}(t) \right] dz_s\, dt$$
$$-V_i(t) h'(t) R^{-1}(t)\, dz_t$$
$$= -V_i(t) h'_i(t) R^{-1}(t)\, dz_t - \sum_k [V_k(t) - 1] a_{ki}(t)\, dt$$
$$= -V_i(t) h'_i(t) R^{-1}(t)\, dz_t - \sum_k V_k(t) a_{ki}(t)\, dt. \qquad (5.11)$$

on using the fact that $\sum_k a_{ki}(t) = 0$. This is the desired equation of evolution for state model F and observation model G.

Now let us consider a state model of type D, but in conjunction with the counting process observation model C. Define

$$\phi_t \overset{f}{=} \int_t^T \ln h(x_s, s)\, dz_s - \int_t^T (h(x_s, s) - 1)\, ds. \qquad (5.12)$$

It is easily established, using the appropriate Ito differentiation rule, see e.g. [11], that the forward equation for $L_{t,T} = \exp(\phi_t)$ is

$$dL_{t,T} \overset{f}{=} L_{t,T}[h^{-1} - 1]\, dz_t + L_{t,T}(h - 1)\, dt \qquad (5.13)$$

and the corresponding backward equation is

$$dL_{t,T} \overset{b}{=} -L_{t,T}[h-1]\,dz_t + L_{t,T}(h-1)\,dt \tag{5.14}$$

so that

$$L_{t,T} \overset{b}{=} \int_t^T L_{s,T}[h-1][dz_s - ds] + 1, \tag{5.15}$$

(with a backward Ito integral applying). Then similar arguments applied to (5.15) as were applied above for observation model G lead to

$$dV_t(x_t) \overset{b}{=} -V_t(x_t)(h-1)[dz_t - dt] - \mathscr{L}^a[V_t(x_t)]\,dt;\ V_T(x_T) = 1 \tag{5.16}$$

where

$$V_t(x_t) = E[L_{t,T} \,|\, Z_{t+}, x_t = x]. \tag{5.17}$$

This equation also appears in [16, Section 5].

Finally, state model F and observation model C lead with similar arguments to

$$dV_i(t) \overset{b}{=} -V_i(t)(h_i - 1)[dz_t - dt] - \sum_k V_k(t)a_{ki}(t)\,dt. \tag{5.18}$$

6. SMOOTHING WITHOUT A REVERSE TIME FILTER OR REVERSE TIME LIKELIHOOD FUNCTION

In [4, 5] the following equation is derived linking the smoothed density $p_s(x_t) = p(x_t | Z_T)$ and the forward filtered density $p_f(x_t) = p(x_t | Z_t)$ for state model D and observation model G:

$$dp_s = \left[\frac{p_s}{p_f}\mathscr{L}(p_f) - p_f\mathscr{L}^a\left(\frac{p_s}{p_f}\right)\right]dt. \tag{6.1}$$

The idea is that the forward filtering density is first found, and then (6.1) is solved backwards in time from T, using the boundary conditions $p_s(x_T) = p_f(x_T)$. Using the ideas of this paper we shall derive this equation and the modifications of it associated with different combinations of state and observation models. Of course, the various densities are all to be interpreted as functions rather than measures.

Let us recall from (3.9) and the definition of $V_t(x)$ in (5.6) that

$$p(x_t = x | Z_T) = p(x_t = x | Z_t)V_t(x_t)\frac{E_x[L_{0,t} | Z_t]}{E_x[L_{0,T} | Z_T]}. \tag{6.2}$$

Accordingly, with p_{su} and p_{fu} denoting unnormalized densities,

$$p_{su}(x_t) = p_{fu}(x_t)V_t(x_t). \tag{6.3}$$

Now in (5.9), we found a backward equation for $V_t(x_t)$. The corresponding forward equation is

$$dV_t(x_t) \overset{f}{=} -\mathscr{L}^a[V_t(x)]\,dt - V_t(x_t)h'(x_t, t)R^{-1}(t)\,dz_t + V_t(x_t)\|h(x_t, t)\|^2_{R^{-1}(t)}\,dt. \tag{6.4}$$

Combining this with the forward Ito equation for $p_{fu}(x_t)$, see (4.8), yields

$$\begin{aligned}
dp_{su}(x_t) &\overset{f}{=} d[p_{fu}(x_t)V_t(x_t)] = d[p_{fu}(x_t)]V_t(x_t) + p_{fu}(x_t)\,dV_t(x_t) + dp_{fu}(x_t)\,dV_t(x_t)\\
&= V_t(x_t)\mathscr{L}[p_{fu}(x_t)]\,dt + V_t(x_t)h'R^{-1}(t)\,dz_t p_{fu}(x_t)\\
&\quad - p_{fu}(x_t)\mathscr{L}^a[V_t(x)]\,dt - p_{fu}(x_t)V_t(x_t)h'R^{-1}(t)\,dz_t\\
&\quad + p_{fu}(x_t)V_t(x_t)\|h\|^2_{R^{-1}(t)}\,dt - V_t(x_t)p_{fu}(x_t)\|h\|^2_{R^{-1}(t)}\,dt\\
&= V_t(x_t)\mathscr{L}[p_{fu}(x_t)]\,dt - p_{fu}(x_t)\mathscr{L}^a[V_t(x_t)]\,dt\\
&= \left\{\frac{p_{su}(x_t)}{p_{fu}(x_t)}\mathscr{L}[p_{fu}(x_t)] - p_{fu}(x_t)\mathscr{L}^a\left[\frac{p_{su}(x_t)}{p_{fu}(x_t)}\right]\right\}dt.
\end{aligned}$$

Now it is immediate that the $\mathscr{L}$ and $\mathscr{L}^a$ operators commute with the normalizing quantity associated with p_{fu}. Accordingly,

$$dp_{su}(x_t) = \left\{\frac{p_{su}(x_t)}{p_f(x_t)}\mathscr{L}[p_f(x_t)] - p_f(x_t)\mathscr{L}^a\left[\frac{p_{su}(x_t)}{p_f(x_t)}\right]\right\}dt. \tag{6.5}$$

It is now easy to obtain the normalized equation, since the integral of the coefficient of dt over x_t space is zero, indicating that the normalizing quantity

$$n_s(t) = \int_{R^n} p_{su}(x_t)\,dx_t,$$

satisfies $dn_s(t) = 0$, i.e. n_s is constant; (6.1) is then immediate.

We turn now to the finite state model F with observation model G. Using (5.11), we have the following forward equation for $V_i(t)$

$-E[L_{t,T} \mid Z_{t}, x_t = i]$:

$$dV_i(t) \overset{f}{=} V_i(t)h_i'(t)R^{-1}(t)\,dz_t \quad \sum_k V_k(t)a_{ki}(t)\,dt + h_i'(t)R^{-1}(t)h_i(t)V_i(t)\,dt. \tag{6.6}$$

Accordingly, with $p_{isu}(t) = p_{ifu}(t)V_i(t)$, (4.25) and (6.6) yield

$$\begin{aligned} dp_{isu}(t) &\overset{f}{=} V_i(t)\,dp_{ifu}(t) + p_{ifu}(t)\,dV_i(t) \quad h_i'(t)R^{-1}(t)h_i(t)V_i(t)p_{ifu}(t)\,dt \\ &= V_i(t)\sum_j a_{ij}(t)p_{jfu}(t)\,dt - \sum_k V_k(t)a_{ki}(t)p_{ifu}(t)\,dt \\ &= \frac{p_{isu}(t)}{p_{ifu}(t)}\left[\sum_j a_{ij}(t)p_{jfu}(t)\right]dt \quad p_{ifu}(t)\left[\sum_k \frac{p_{ksu}(t)}{p_{kfu}(t)}a_{ki}(t)\right]dt. \end{aligned}$$

Since all unnormalized filtered probabilities occur as ratios, we have

$$dp_{isu}(t) = \frac{p_{isu}(t)}{p_{if}(t)}\left[\sum_j a_{ij}(t)p_{jf}(t)\right]dt - p_{if}(t)\left[\sum_j \frac{p_{jsu}(t)}{p_{jf}(t)}a_{ji}(t)\right]dt.$$

Now observe that this equation implies easily that

$$\sum_i dp_{isu}(t) = 0.$$

Thus the normalization on $p_{isu}(t)$ is independent of time. Accordingly,

$$dp_{is}(t) = \frac{p_{is}(t)}{p_{if}(t)}\left[\sum_j a_{ij}(t)p_{jf}(t)\right]dt \quad p_{if}(t)\left[\sum_j \frac{p_{js}(t)}{p_{jf}(t)}a_{ji}(t)\right]dt. \tag{6.7}$$

Of course the initialization for the equation is provided by $p_{is}(T) = p_{if}(T)$ and the equation is solved backwards in time. This equation has been obtained by different arguments in [6].

It is interesting to note that we can rewrite (6.7), with p_s denoting the vector of p_{is},

$$dp_s = A^s p_s\,dt, \tag{6.8}$$

where A^s is a stochastic matrix and A^s is defined by

$$a_{ii}^s = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}p_{jf}}{p_{if}} \tag{6.9a}$$

$$a_{ij}^s = \frac{p_{jf}a_{ji}}{p_{if}}. \tag{6.9b}$$

Now consider a state model D with observation process C. Eq. (6.3) is still relevant. The forward equation for $V_t(x_t)$ is, from (5.16),

$$dV_t(x_t) \overset{f}{=} V_t(x_t)[h^{-1} - 1]\,dz_t + V_t(x_t)[h - 1]\,dt - \mathscr{L}^a[V_t(x_t)]\,dt. \tag{6.10}$$

[The argument is just like that for connecting (5.13) and (5.14).] Then combining this with (4.30), for $p_{fu}(x_t)$, we obtain

$$\begin{aligned} dp_{su}(x_t) &\overset{f}{=} V_t(x_t)\,dp_{fu}(x_t) + p_{fu}(x_t) + dV_t(x_t)\,dp_{fu}(x_t) \\ &= V_t(x_t)\mathscr{L}(p_{fu})\,dt + V_t(x_t)(h - 1)(dz - dt)p_{fu} \\ &\quad + p_{fu}V_t(x_t)(h^{-1} - 1)\,dz + p_{fu}V_t(x_t)(h - 1)\,dt \\ &\quad - p_{fu}\mathscr{L}^a[V_t(x_t)]\,dt + p_{fu}V_t(x_t)(h - 1)(h^{-1} - 1)\,dz \\ &= V_t(x_t)\mathscr{L}(p_{fu}(x_t))\,dt - p_{fu}(x_t)\mathscr{L}^a(V_t(x_t))\,dt, \end{aligned} \tag{6.11}$$

and from this we immediately obtain the same Eq. (6.5) as was found to hold for observation model G. In other words, Eq. (6.5) holds for either observation model G or P in conjunction with state model D.

In the same vein, similar calculations show that Eq. (6.7) holds not only for state model F and observations G as shown above, but also for state model F and observations C: one uses (4.25) for p_{ifu} and a forward time version of (5.18) to obtain (6.7).

7. FINITE DIMENSIONAL SMOOTHERS

We have already indicated in the last two sections one class of finite dimensional smoothers those associated with a finite state model (type F). The forward filter, reverse filter, reverse likelihood function, and *a priori* probability computations are all finite-dimensional. In this section, we note some other possibilities of finite-dimensional filters.

Linear-Gaussian problems

This case is well known, and the three distinct approaches advanced here all have their specializations which have appeared in the literature. We comment, however, that the two Eqs. (3.1) and (3.2) do not seem to have been recognized as such in the linear-Gaussian case. Let m_u, m_f, m_r, m_s denote the unfiltered, forward-filtered, reverse-filtered, and smoothed means. Let Σ_u, Σ_f, Σ_r, Σ_s denote the corresponding (error) covariances. Then it is a trivial observation from (3.1), knowing the densities are Gaussian, to obtain

$$\Sigma_s^{-1}=\Sigma_f^{-1}+\Sigma_r^{-1}-\Sigma_u^{-1} \qquad (7.1)$$

and

$$m_s=\Sigma_s[\Sigma_f^{-1}m_f+\Sigma_r^{-1}m_r-\Sigma_u^{-1}m_u]. \qquad (7.2)$$

These equations appear to have first been derived in [22 23]. These references also include equations for the update of the mean m_r and covariance Σ_r associated with the reverse filter, and thus capture the approach of Section 4.

As described also in [22 23], the Fraser Potter smoothing formula [23, 24] can be thought of as combining a forward filtered estimate and a reverse maximum likelihood estimate. Thus (3.2), or (3.9), is the key formula applicable here. One can show in the linear Gaussian case that $E_x[L_{t,T}\,|\,Z_{t,T}, x_t=x]$ is proportional to

$$\exp-\tfrac{1}{2}[(x-m_L)'\Sigma_L^{-1}(x-m_L)]$$

where

$$\Sigma_L^{-1}=\Sigma_r^{-1}-\Sigma_u^{-1} \qquad (7.3)$$

and

$$\Sigma_L^{-1}m_L=\Sigma_r^{-1}m_r-\Sigma_u^{-1}m_u. \qquad (7.4)$$

Again, formulas are given in the references for the update of M_L and Σ_L, and so the ideas of Section 5 are captured.

The formula (6.1) also has its counterpart, see for example [25, Section 7.5]; equations for the smoothed mean and error covariance are given in terms of a priori information and the forward filtered mean and error covariance, with no use of the measurement process, and the equations are solved backwards in time from T.

Problems with space-time point-process observations

In [26], an extension of the standard linear Gaussian problem is considered. This consists in providing an additional observation process which is a space-time point process defined on $[0,\infty)\times R^m$ as follows.

Each point occurrence is identified by a temporal coordinate $t\in[0,\infty)$ and a spatial coordinate $r\in R^m$. Let τ, A be Borel sets in $[0,\infty)$, R^m and denote by $N(\tau\times A)$ the number of points occurring in $\tau\times A$, with $N_t\overset{\Delta}{=}N([0,t)\times R^m)$ being the number of points occurring before t at any location; N is taken to be a doubly stochastic Poisson counting process with intensity μ, where the stochastic processes μ and N are independent of the state process and the linear-in-the-state observation process, and μ_t is a.s. positive. Given that N has a jump at t, the spatial location r of the point is an m-dimensional Gaussian random vector with mean $H(t)x_t$ and known positive definite covariance $S(t)$, where $H(\cdot)$ is known. Given N_s and x_s for $s\geqq 0$, the spatial locations are independent random vectors that are independent of all other entities. Thus the space-time process can be thought of as having an intensity

$$\lambda_t(r,x_t,\mu_t)=\mu_t\gamma_t(r,x_t) \qquad (7.5)$$

that separates into the product of a temporal component μ_t that underlies the Poisson counting process N and a spatial component

$$\gamma_t(r,x_t)=(2\pi)^{-m/2}[\det S(t)]^{-1/2}\exp\{-\tfrac{1}{2}\|r-H(t)x_t\|^2_{S^{-1}(t)}\}. \qquad (7.6)$$

The main result of [26] is that the conditional density of x_t given the two observation processes up until time t is (conditionally) Gaussian; both the conditional mean and the conditional covariance satisfy nonlinear but finite-dimensional stochastic differential equations driven by the observation processes.

Now, as noted in Section 4, a reverse time model can be found in general for the linear-Gaussian state process and the linear-Gaussian observation process. The reverse models are linear and Gaussian. The point process has an unchanged reverse model. It follows that the reverse filtered density is also conditionally Gaussian (and so, by the fundamental formula (3.1), is the smoothed density). Further, the reverse conditional mean and covariance can be found in the same manner as the corresponding forward quantities.

8. CONCLUSIONS

The ideas of the paper fall into four divisions. First are the pair of equations relating the smoothed density (or probability, as the case may be) to the forward filtered, reverse filtered, and unfiltered densities or, alternatively, to the forward filtered density and a reverse-time likelihood ratio. The second division is the development of reverse-time filtered equations, using recent results on the construction of reverse time processes. Third is the development of equations for the evolution of the reverse-time likelihood ratio, and last is the development of equations that relate smoothed and filtered densities and are undriven by the measurement.

Though we have not discussed the point, it is clear that there are reverse-time equivalents of the last two ideas, which would involve a forward-evolving likelihood ratio to be used in conjunction with a reverse filtered density to get a smoothed density, and an equation relating smoothed and reverse filtered densities, undriven by the measurements.

The results have all been presented with certain independence assumptions between system state and measurement noise; for example, with state model *D* and measurement model *G*, the processes w_t and v_t have been assumed independent. Assumptions like this can doubtlessly be relaxed.

It would also be possible to develop parallels of a number of the ideas in discrete time. A quite different development could lie in the generation of bounds on the smoothed errors, such as can be done in the filtering case. Indeed, the key formula relating smoothed, forward filtered, reverse filtered, and unfiltered quantities may be of great help here.

References

[1] P Whittle, *Reversibility and Acyclicity*, pp 217 224, in *Perspectives in Probability and Structures*, ed. J. Gani, Academic Press, London, 1975.

[2] B D. O Anderson and T Kailath, Forwards and backwards models for finite-state Markov processes, *Advances in Applied Probability* **11** (1979), 118 133.

[3] B D. O. Anderson, Reverse-time diffusion equation models, *Stochastic Processes and their Applications* **12** (1982), 313 326.

[4] C T Leondes, J B Peller and E B. Stear, Nonlinear smoothing theory, *IEEE Trans. Syst Science Cyber.* **SSC-6** (1970), 63 71

[5] B D O. Anderson, Fixed interval smoothing for nonlinear continuous-time systems, *Information and Control* **20, No. 3.** (April 1972), 294 300

[6] R S. Liptser and A N Shiryaev, *Statistics of Random Processes I*, Springer-Verlag, Berlin, 1977.

[7] E Pardoux, Stochastic partial differential equations and filtering of diffusion processes, *Stochastics* **3** (1979), 127 167.

[8] E Pardoux, Equations du filtrage nonlinéaire, de la prédiction et du lissage, *Stochastics* **6** (1982), 193 231.

[9] P Frost and T Kailath, An innovative approach to least-square estimation Part III: Nonlinear estimation in white Gaussian noise, *IEEE Trans. Auto Control* **AC-16** (June 1971), 217 226

[10] D L Snyder, Smoothing for doubly stochastic Poisson processes, *IEEE Trans Info Theory* **IT-18** (1972), 558 562

[11] K. P. Dunn and I. B. Rhodes, A generalized representation theorem with applications to estimation and hypothesis testing, *Proc. Eleventh Allerton Conference on Circuit and System Theory*, (1973), 304 314.

[12] D. Clements and B. D. O. Anderson, A nonlinear fixed-lag smoother for finite-state Markov processes, *IEEE Trans. Info. Theory* **IT-21** (1975), 446 452.

[13] E Wong, *Stochastic Processes in Information and Dynamical Systems*, McGraw Hill, New York, 1971.

[14] J H. VanSchuppen, Stochastic filtering theory: A discussion of concepts, methods, and results, in *Stochastic Control Theory and Stochastic Differential Systems*, M. Kolhmann and W. Vogel, eds., New York, Springer-Verlag, 1979.

[15] R S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*, Wiley-Interscience, New York, 1968.

[16] E. Pardoux, Non-linear filtering, prediction and smoothing, in M. Hazewinkel and J. C. Willems (eds.), *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, Reidel, Dordrecht, 1981.

[17] A. Lindquist and G. Picci, On the stochastic realization process problem, *SIAM J. Control Opt.* **17** (May 1979), 365 389.

[18] M. Zakai, On the optimal filtering of diffusion processes, *Zeitschrift fur Wahrscheinlichkeitstheorie der Verwandten Gebiete*, **11** (1969), 230 243.

[19] M. H. A. Davis and S. I. Marcus, An introduction to nonlinear filtering, in M. Hazelwinkel and J. C. Willems, eds., *Stochastic Systems; The Mathematics of Filtering and Identification*, NATO Advanced Study Institute Series, Reidel, Dordrecht, 1980

[20] W. M. Wonham, Some applications of stochastic differential equations to optimal nonlinear filtering, *SIAM J. Control* **2** (1965), 347 369.

[21] D. L. Snyder, *Random Point Processes*, Wiley, New York, 1975.

[22] J. E. Wall, A. S. Willsky and N. R. Sandell, On the fixed-interval smoothing problem, *IEEE Trans. Auto Control* to appear.

[23] D. C. Fraser, *A New Technique for the Optimal Smoothing of Data*, Sc.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1967.

[24] D. C. Fraser and J. E. Potter, The optimum linear smoother as a combination of two optimum linear filters, *IEEE Trans. Auto. Control* **AC-14** (August 1969), 387 390.

[25] J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, McGraw Hill, New York, p. 261, 1969.

[26] I. B. Rhodes and D. L. Snyder, Estimation and control performance for space-time point-process observations, *IEEE Trans Auto. Control* **AC-22** (June 1977), 338 346

APPENDIX A—EXTENDED DERIVATION OF EQ. (4.10)

We start with the backward Eq. (4.5), which can be written as

$$x_t - x_{t\ dt} \overset{b}{=} f^r(x_t, t)\,dt + g(x_t, t)[v_t^r - v_{t\ dt}^r].$$

To prevent confusion below, let us rewrite this as

$$x_t - x_{t\ \Delta} \overset{b}{=} f^r(x_t, t)\Delta + g(x_t, t)[v_t^r - v_{t\ \Delta}^r]$$

where it is understood that Δ is a vanishingly small positive quantity.

Now let

$$\tilde{x}_t = x_{T-t}$$

$$\tilde{v}_t^r = v_{T-t}^r$$

$$\tilde{f}^r(\tilde{x}_t, t) = f^r(\tilde{x}_t, T-t) = f^r(x_{T-t}, T-t)$$

$$\tilde{g}^r(\tilde{x}_t, t) = g^r(\tilde{x}_t, T-t) = g^r(x_{T-t}, T-t).$$

Then

$$\tilde{x}_{t+\Delta} - \tilde{x}_t = x_{T-(t+\Delta)} - x_{T-t}$$

$$= -[x_{(T-t)} - x_{(T-t)-\Delta}]$$

$$= -f^r(x_{T-t}, t-t)\Delta - g(x_{T-t}, T-t)[v_{T-t}^r - v_{T-t-\Delta}^r]$$

$$= -\tilde{f}^r(\tilde{x}_t, t)\Delta - \tilde{g}(\tilde{x}_t, t)[\tilde{v}_t^r - \tilde{v}_{t+\Delta}^r]$$

$$\overset{f}{=} -\tilde{f}^r(\tilde{x}_t, t)\Delta + \tilde{g}(\tilde{x}_t, t)[\tilde{v}_{t+\Delta}^r - \tilde{v}_t^r]. \qquad \text{(A.1)}$$

This of course is a forward Ito equation, since Δ is still a vanishingly small *positive* quantity.

To get an associated observation equation, first of all write (4.7b) as

$$z_t - z_{t-\Delta} \overset{b}{=} h(x_x, t)\Delta + M(t)[w_t - w_{t-\Delta}].$$

Set

$$\tilde{z}_t = z_{T-t} \qquad \tilde{w}_t = w_{T-t}$$

$$\tilde{h}(\tilde{x}_t, t) = h(\tilde{x}_t, T-t) = h(x_{T-t}, T-t)$$

$$\tilde{M}(t) = M(T-t).$$

Then

$$\tilde{z}_{t+\Delta} - \tilde{z}_t = -[z_{(T-t)} - z_{(T-t)-\Delta}]$$

$$= -h(x_{T-t}, T-t)\Delta - M(T-t)[w_{T-t} - w_{T-t-\Delta}]$$

$$\overset{f}{=} \tilde{h}(\tilde{x}_t, t)\Delta + \tilde{M}(t)[\tilde{w}_{t+\Delta} - \tilde{w}_t]. \qquad \text{(A.2)}$$

Now, (A.1) and (A.2) are both forward equations to which the forward filtering equation can be applied. Thus, see (4.8)

$$p_u(\tilde{x}_{t+\Delta} = \alpha \mid \tilde{z}_s, 0 \leqq s \leqq t+\Delta) - p_u(\tilde{x}_t = \alpha \mid \tilde{z}_s, 0 \leqq s \leqq t)$$

$$\overset{f}{=} \mathscr{L}[p_u(\tilde{x}_t = \alpha \mid \tilde{Z}_t)]\Delta - p_u(\tilde{x}_t = \alpha \mid \tilde{Z}_t)\hat{\tilde{h}}_t' R^{-1}(\tilde{z}_{t+\Delta} - \tilde{z}_t). \qquad \text{(A.3)}$$

We must now convert this to a backward equation. We have at once:

$$p_u(\tilde{x}_{T-t+\Delta} = \alpha \mid \tilde{z}_s, 0 \leqq s \leqq T-t+\Delta) - p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{z}_s, 0 \leqq s \leqq T-t)$$

$$\overset{f}{=} \mathscr{L}[p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{Z}_{T-t})]\Delta$$

$$- p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{Z}_{T-t})\hat{\tilde{h}}_{T-t}' R^{-1}(\tilde{z}_{T-t+\Delta} - \tilde{z}_{T-t}). \qquad \text{(A.4)}$$

Here

$$\mathscr{L}[p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{Z}_{T-t})] = -\sum_i \frac{\partial}{\partial \alpha^i}[-\tilde{f}^{ri}(\alpha, T-t)p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{Z}_{T-t})]$$

$$+ \tfrac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial \alpha^i \partial \alpha^j}\{[\tilde{g}(\alpha, T-t)g'(\alpha, T-t)]^{ij} p_u(\tilde{x}_{T-t} = \alpha \mid \tilde{Z}_{T-t})\}$$

$$= -\sum_i \frac{\partial}{\partial \alpha^i}\{-f^{ri}(\alpha, t)p_u(x_t = \alpha \mid z_s, t \leqq s \leqq T\} + \tfrac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial \alpha^i \partial \alpha^j}$$

$$\times \{[g(\alpha, t)g'(\alpha, t)]^{ij} p_u(x_t = \alpha \mid z_s, t \leqq s \leqq T)\} = \mathscr{L}_r[p_{ru}]. \qquad \text{(A.5)}$$

where $\mathscr{L}_r$ is defined like $\mathscr{L}$, save that $-f^r$ replaces f, and p_{ru} is shorthand for $p_u(x_t = \alpha \mid z_s, t \leqq s \leqq T)$.

Using (A.5) in (A.4), we have

$$p_u(x_{t-\Delta} = \alpha \mid z_s, t-\Delta \leqq s \leqq T) - p_u(x_t = \alpha \mid z_s, t \leqq s \leqq T)$$

$$\overset{b}{=} \mathscr{L}_r[p_{ru}]\Delta - p_{ru}\hat{h}' R^{-1}(z_{t-\Delta} - z_t)$$

or

$$dp_{ru} \overset{b}{=} -\mathscr{L}_r[p_{ru}]\,dt - p_{ru}\hat{h}' R^{-1}\,dz$$

as desired.

Reprint of Summary:

"Near Disturbance Localization using Second-Order Modes", J. M. Saniuk and I. B. Rhodes, 23rd Annual Allerton Conference on Control, Communications and Computing, University of Illinois, October 1985.

NEAR-DISTURBANCE LOCALIZATION USING SECOND-ORDER MODES

J.M. Saniuk and I.B. Rhodes
Dept. of Electrical & Computer Engineering
University of California
Santa Barbara, CA 93106

SUMMARY

The product of the observability and reachability gramians of a linear system, denoted MW, is a matrix whose eigenvalues are non-negative and are invariant under coordinate transformations in the state space [4]. The *second-order modes* of the system-- the square roots of these eigenvalues-- can be interpreted as measures of the mean-square energy throughput in appropriately defined channels [3,6 and others]. The sum of the eigenvalues also arises in a natural way as a continuous-valued measure of interaction for purposes of decoupling, or nearly decoupling, a system with several input and output channels: define measures of reachability and observability, respectively, on the state space by R(x)=$\frac{1}{2}x^T Wx$ and O(x) = $\frac{1}{2}x^T Mx$ [2]. Then let a measure of interaction between the system inputs and outputs be defined as

$$I = \underset{R^*(x)=1}{E} O(x)$$

where $R^*(x)$ denotes the conjugate functional $\frac{1}{2}x^T W^{-1}x$, and where the expectation is with respect to the uniform density on $\{R^*(x) = 1\}$. It is readily seen (as in [7]) that I = 1/n tr MW. For convenience, we shall use the measure tr MW.

We present results for near-disturbance localization in discrete-time systems based on satisfaction of a first-order necessary condition. Consider a system

$$x_{k+1} = A x_k + B u_k + D v_k, \tag{1a}$$

$$y_k = C x_k \tag{1b}$$

and consider u_k as a local control input, v_k a disturbance. Define the problem of near-disturbance localization by state feedback as: find F = arg min tr $M_F W_{D.F}$, where $(\cdot)_F$ signifies that the system matrix is (A+BF), and where $W_{D.F}$ denotes the reachability matrix of the pair ((A+BF),D). Consider the case in which A_F is stable. For the infinite-horizon problem in which M_F, $W_{D.F}$ are steady-state gramians, the gradient matrix of tr $M_F W_{D.F}$ with respect to the feedback gain matrix F has a concise closed-form expression:

$$\frac{\partial}{\partial F} tr\ M_F W_{D.F} = 2B^T (M_F A_F \Gamma_{D.F} + \Lambda_F A_F W_{D.F})$$

where Γ and Λ are *weighted gramians* defined by the steady-state discrete-time Lyapunov equations

$$\Gamma_{D.F} = A_F \Gamma_{D.F} A_F^T + W_{D.F}$$
$$\Lambda_F = A_F^T \Lambda_F A_F + M_F$$

We wish to find F such that $\frac{\partial}{\partial F}$tr $M_F W_{D.F}$=0. If there are at least as many control inputs as outputs, with B having full (column) rank, this condition has a solution that is independent of D. Let p be the number of outputs. Consider first the case in which the first nonzero Markov parameter of the local system has full rank: $\rho(CB) = \cdots = \rho(CA^{\delta-2}B) = 0$ and $\rho(CA^{\delta-1}B)$=p. (This is always true of single-output systems.) Then it is easily seen that $F^* = -(CA^{\delta-1}B)^+ CA^{\delta}$, where $(CA^{\delta-1}B)^+$ is any right inverse for $CA^{\delta-1}B$. This F^* has the property that $F^* \in F(V^*)$, and the map induced by $(A+BF^*)$ on $\mathbf{X}/\mathbf{V}^*$ has all its eigenvalues at the origin; this property is a characterization of F^* for single-input, single-output systems. Furthermore, if the system is minimum-phase, F^* is the same as the optimal F obtained by the "cheap control" technique of finding

$$F^o = \lim_{\varepsilon \to 0} F_\varepsilon^o ,$$

where

$$u_k = F_\varepsilon^o x_k$$
$$= arg\ min \left\{ \sum_{k=0}^{\infty} x_k^T C^T C x_k + \varepsilon u_k^T u_k \right\}.$$

This is equivalent to finding F^o = arg min. tr $C\, W_{D,F}\, C^T$ in the case of a minimum-phase system; the gradient of this latter interaction measure is given by

$$\frac{\partial}{\partial F} tr\; C\, W_{D,F}\, C^T = 2B^T(\, M_F\, A_F\, W_{D,F}\,).$$

$F^* = -(CA^{\delta-1}B)^+ CA^{\delta}$ satisfies both first-order necessary conditions; in fact, it satisfies the stronger condition

$$(C(A+BF^*)^{k-1}B)^T(C(A+BF^*)^k) = 0, \;\; k=1,2,\ldots, \tag{2}$$

which ensures that both gradients vanish term-by-term.

In the nonuniform rank case in which m $\geq$p>1 and CB =...=C$A^{\delta-2}$B =0, C$A^{\delta-1}$B $\neq$0 and $\rho(CA^{\delta-1}B)<$ p, one plausible extension of the above results is that the optimal F^* be characterized by the condition of eqn. (2), and be described by the property that 1) $F^* \in F(V^*)$; 2) the map induced by $(A+BF^*)$ on $\mathbf{X}/\mathbf{V}^*$ has all eigenvalues at the origin; and 3) $F^* = -(B^T P B)^{-1}\; B^T P A$, where P is a (not necessarily stabilizing) solution to the Riccati equation $P = A^T P A + C^T C - A^T P B (B^T P B)^{-1} B^T P A$. The simulations performed so far support this conjecture.

References.

[1] W.M. Wonham, *Linear Multivariable Control: A Geometric Approach* (2nd ed.), New York: Springer-Verlag, 1979.

[2] I.B. Rhodes, "Some quantitaitve measures of controllability and observability and their implications," in *Proc. 8th IFAC Congress*, Kyoto, Japan, 1981, pp.30-35.

[3] B.C. Moore, "Principal component analysis in linear systems: controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.* vol. AC-

26, pp. 17-32, February 1981.

[4] C.T. Mullis and R.A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-562, September 1976.

[5] J.L. Willems, "Disturbance isolation in linear feedback systems," *Int. J. Systems Science*, vol. 6, pp. 233-238, 1975.

[6] E. A. Jonckheere and L.M.Silverman, "A new set of invariants for linear systems -- application to reduced order compensator design," *IEEE Trans. Automat. Contr.*, vol. AC-28, no.10, Oct. 1983, pp.953-964.

[7] M. Athans and E.Tse, "A Direct Derivation of the Optimal Linear Filter Using the Maximum Principle", *IEEE Trans. Automat. Contr.*, vol. AC-12,no.6,Dec.1967, pp.690-698.

[8] H.Kwakernaak and R.Sivan, *Linear Optimal Control Systems*. New York: Wiley, 1972.

[9] H.Kwakernaak and R.Sivan, "The maximally achievable accuracy of linear optimal regulators and linear optimal filters," *IEEE Trans. Automat. Contr.*, vol. AC-17,no.1,Febr. 1972, pp.79-86.

Reprint of Summary:

"Uniqueness of the Minimal Dynamic Cover and the Associated Solution to Sylvester's Equation", A. F. Assal and I. B. Rhodes, 24th Annual Allerton Conference on Control, Communications and Computing, University of Illinois, October 1986.

UNIQUENESS OF THE MINIMAL DYNAMIC COVER AND THE ASSOCIATED SOLUTION TO SYLVESTER'S EQUATION

A F ASSAL AND I B RHODES
Electrical and Computer Engineering Department
University of California at Santa Barbara
Santa Barbara, Ca 93106

1. INTRODUCTION

In this paper we give necessary and sufficient conditions for the uniqueness of the minimal dynamic cover [1] for a *prescribed* subspace ***K***. Also given are necessary and sufficient conditions for the *uniqueness (modulo a similarity transformation)* of the matrix F that satisfies, for a given pair (A',C'), Sylvester's equation $TA - FT = GC$ *for some* matrix G with appropriate dimension, where the range of T' is a minimal cover for ***K***. *These* necessary and sufficient conditions are in terms of the obsevability indices of the augmented pair ((C' | K')',A) *where the* columns of K span the subspace***K***.

For a given pair (C, A) a **dynamic cover** *of a* prescibed subspace, ***K***, is *defined* in [1] as any (A',C') invariant subspace [2] that, *together* with ***C***, the range of C', contains ***K*** *Hence, if* τ_{dc} is a *cover for **K** it satisfies the equations*

$$A'\tau_{dc} \subset \tau_{dc} + C \qquad (1)$$

$$K \subset \tau_{dc} + C \qquad (2)$$

A cover of a given subspace is *minimal* if there is no cover with smaller dimension among all the covers for of this subspace

Emre et al [3] gave necessary and *sufficient* conditions for the uniqueness of the minimal cover of a given subspace The procedure involved solving a succession of subproblems each identical to the original and having a lower dimension than the previous one The solution to any subproblem was obtained by solving the next subproblem *in the sequence* Emre also showed that the problem of finding a *minimal dynamic cover is equivalent to the problem of finding* a minimal realization i e matrices (C, A, B), of a transfer function matrix $G(z) = A^{-1}(z)B(z)$.

2. PRELIMINARIES

1. Consider the augmented system (K_c, A), where K_c=(C' | K')', where the range of K' is the prescribed subspace ***K***, and compute the observability indices [1,4] associated with the *augmented system*

2. *Let* m_i, *for* i=1, , *m*, and n_j, for j=1, , q, be the observability indices associated, respectively, with the rows of C and the rows of K

3. Assuming that (C, A) is *completely observable then*

$$\sum_{i=1}^{m} m_i + \sum_{j=1}^{q} n_j = n$$

and the dimension of the *minimal cover* of ***K*** is

$$\sum_{i=1}^{q} n_i$$

4. A cover for ***K*** satisfies eqns (1) and (2) which are, as shown in [5], equivalent, respectively, to

$$TA - FT = GC \qquad \textit{(Sylvester's Equation)} \qquad (1a)$$
$$K = MT + NC \qquad (1b)$$

Hence if the columns of T' span a cover of ***K*** then such a T' satisfies *the two latter equations* for some F, G, M and N

3 RESULTS

Uniqueness of the Minimal Cover

In this Section, necessary and sufficient conditions *for uniqueness* of the minimal cover for ***K*** are given in terms of the observability indices associated with the rows of K and C If the columns of T' span a cover of ***K***, then T satisfies Sylvester's equation

$$TA - FT = GC \qquad (1a)$$

for some F and G Uniqueness of the cover implies uniqueness of T (modulo some row operations)

Theorem 1

*The minimal cover of a given subspace **K** is unique iff all observability indices associated with the rows of C are greater than all observability indices associated with the rows of K, where the the columns of K' span the subspace **K**.*

Uniqueness of the F matrix that Satisfies Sylvester's Equation Under specified Conditions

The following theorem gives necessary and sufficient conditions for the uniqueness of the matrix F that satisfies the equation $TA - FT = GC$, where $\tau_i = R(T_i)'$ spans a *minimal cover* for some prescribed ***K***.

Theorem 2

*The matrix F that satisfies the equation TA - FT = GC, where the columns of T' form a basis for a minimal cover of a prescribed subspace **K**, is unique (modulo a similarity transformation), iff*

I. All m_i, for i =1,.., m, *are greater than* largest n_j, for j=1, ..., q or,

II. For all m_i, for i =1,...., m, that are less than or equal to the largest n_j, for j=1,.., q, $(c_i)'A^{m_i}$ *does not have components* along the rows of Q i.e

$$(c_i)'A^{m_i} = (v_i)'C_a$$

where the columns of Q' span any cover *for **K*** and those of (Q'|C_a') span the whole space, and C_a=[(M₁)' .....(Mₘ)']' and $(M_i)' = [c_i \ldots\ldots (A^{m_i-1})' c_i]$, ∀ i =1,...., m and c_i is the i^{th} column of C'

Remarks

1. Both theorems are proven *by solving Sylvester's equation and showing that the uniqueness* of T (modulo some row operations) in Theorem 1, and that *the uniqueness* of F (modulo a similarity transformation) in Theorem 2, holds iff the conditions given in the Theorems are satisfied Complete proofs are given in [6,7]

2 We conclude from Theorems 1 and 2 that uniqueness of the *minimal cover implies uniqueness* of F (*modulo a similarity transformation*) that *satisfies the equation TA - FT = GC, where the columns* of T' span the unique minimal cover On the other hand the converse is not true The uniqueness of F (modulo a similarity transformation) of dimension equal to that of the *minimal cover for the subspace **K*** does not imply uniqueness of the minimal cover for K

REFERENCES

[1] W M Wonham and A S Morse, "Feedback Invariants of Linear Multivariable systems," Automatica, Vol 8, No 1, pp 93-100, Jan 1972

[2] W M Wonham, "Linear Multivariable Control - a Geometric Approach," (New York Springer-Verlag), 1979

[3] E Emre, L M Silverman and K Glover, "Generalized Dynamic Cover for Linear Systems with Applications to Deterministic Identification and Realization Problems," IEEE Trans on Automatic Control, Vol AC-22, No 1, pp 26-35, February 1977

[4] H *Kimura*, "*Geometric Structure* of Observers for Linear Feedback Control Laws," IEEE Trans on Automatic Control, Vol AC-22, No 5, pp 846 -855, October 1977

[5] W M Wonham, "Dynamic Observers- *Geometric Theory*," IEEE Trans on Automatic Control, Vol AC-15, No 2, pp 258-259, April 1970

[6] A Assal, "Issues in the Design of Reduced Order Observers," Ph D Dissertation, University of California at Santa Barbara,May 1986

[7] A Assal and I B Rhodes,"Minimum Order Observers A New Algorithm," to be submitted to the IEEE Trans on Automatic Control

Reprint of Technical Note:

"A Matrix Inequality Associated with Bounds on the Solutions of Algebraic Riccati and Lyapunov Equations", Joan M. Saniuk and Ian B. Rhodes, *IEEE Transactions on Automatic Control*, **AC-32**, No. 8, August, 1987.

A Matrix Inequality Associated with Bounds on Solutions of Algebraic Riccati and Lyapunov Equations

JOAN M. SANIUK AND IAN B. RHODES

Abstract—A new proof is presented for the inequality $\operatorname{tr}(XY) \leq \|X\|_2 \cdot \operatorname{tr} Y$. This argument is valid under the condition that Y be real symmetric nonnegative definite; X may be any square matrix.

Introduction

Much work has been done in recent years to establish bounds on the eigenvalues, and, in particular, on the spectral norm, of solutions to the algebraic matrix Riccati equations and the Lyapunov equations of control and estimation theory. The derivations in, e.g., [2]–[5] have either used or implied the fact that, for $X, Y \in R^{n \times n}$ with both $X, Y \geq 0$,[1]

$$\operatorname{tr}(XY) \leq \|X\|_2 \cdot \operatorname{tr}(Y) \tag{1}$$

where, following [1], $\|\cdot\|_2$ denotes the spectral norm or largest singular value. In [2]–[6], this inequality was only applied to matrices X, Y that were guaranteed to be symmetric and either nonnegative definite or positive definite. A more recent work on eigenvalue bounds [7] contains the related result for $Y = Y' \geq 0$ and $X = X'$

$$\lambda_{\min}(X) \operatorname{tr}(Y) \leq \operatorname{tr}(XY) \leq \lambda_{\max}(X) \operatorname{tr}(Y)$$

which implies (1). However, (1) holds whenever *at least* the matrix Y is symmetric and nonnegative definite; the other can be an arbitrary real-valued square matrix. This more general result has not, to the authors' knowledge, previously appeared in the control literature.

Results

Theorem: Let $X, Y \in R^{n \times n}$ with Y symmetric and nonnegative definite. Then

$$\operatorname{tr}(XY) \leq |\operatorname{tr}(XY)| \leq \|X\|_2 \cdot \operatorname{tr}(Y). \tag{2}$$

Proof: Obviously we need only prove the second inequality. Since Y is symmetric and nonnegative definite, it can be decomposed as $Y = QQ'$, where $Q \in R^{n \times m}$ has full rank $m \leq n$. Write

$$Q = [q_1 | q_2 | q_3 | \cdots | q_m].$$

Then

$$\operatorname{tr} XY = \operatorname{tr} X(QQ') = \operatorname{tr} X[q_1 | q_2 | \cdots | q_m] \begin{bmatrix} q_1' \\ \cdots\cdots \\ q_2' \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ q_m' \end{bmatrix}$$

$$= \operatorname{tr} \sum_{j=1}^{m} X(q_j q_j')$$

and clearly the trace and finite sum can be interchanged to give

$$\operatorname{tr} XY = \sum_{j=1}^{m} \operatorname{tr} X(q_j q_j')$$

$$= \sum_{j=1}^{m} \operatorname{tr}(q_j' X q_j)$$

or

$$\operatorname{tr}(XY) = \sum_{j=1}^{m} q_j' X q_j. \tag{3}$$

Thus,

$$|\operatorname{tr}(XY)| \leq \sum_{j=1}^{m} |q_j' X q_j|$$

$$\leq \sum_{j=1}^{m} \|q_j\| \cdot \|Xq_j\|$$

$$\leq \sum_{j=1}^{m} \|X\|_2 \cdot \|q_j\|^2 = \|X\|_2 \cdot \sum_{j=1}^{m} \|q_j\|^2 \tag{4}$$

by the Cauchy–Schwarz inequality and the definition of the spectral norm. But

$$\sum_{j=1}^{m} \|q_j\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} |q_{ij}|^2$$

$$\triangleq \|Q\|_F^2;$$

and if we recall that $\|Q\|_F^2 = \operatorname{tr} Q'Q = \operatorname{tr} QQ'$, (4) becomes

$$|\operatorname{tr}(XY)| \leq \|X\|_2 \cdot \operatorname{tr} QQ' = \|X\|_2 \cdot \operatorname{tr} Y.$$

References

[1] B. Noble and J. W. Daniel, *Applied Linear Algebra*. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[2] R. V. Patel and M. Toda, "On norm bounds for algebraic Riccati and Lyapunov equations," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 87–88, 1978.

[3] V. R. Karanam, "Lower bounds on the solution of Lyapunov matrix and algebraic Riccati equations," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 1288–1290, 1981.

[4] V. R. Karanam, "A note on eigenvalue bounds in the algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 109–111, 1983.

[5] T. Mori, N. Fukuma, and M. Kuwahara, "On the discrete Lyapunov matrix equation," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 463–464, 1982.

[6] T. H. Kerr, "Three important matrix inequalities currently impacting control and estimation applications," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 1110–1111, 1978.

Manuscript received August 21, 1986; revised March 16, 1987.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

IEEE Log Number 8715227.

[1] [6] assumes both $X, Y > 0$.

7] S.-D. Wang, T.-S. Kuo, and C.-F. Hsu, "Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equations," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 654–656, 1986.

8] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1. New York: Chelsea, 1957.

Preprint of Paper:

“Decentralized Dynamic Decision Making”, H. R. Hashemi and Ian B. Rhodes, Proceedings of the 1987 IEEE Conference on Decision and Control, Los Angeles, California, December 1987, pp. 1836-1841.

FA8 · 11:15

Proceedings of the 26th Conference
on Decision and Control
Los Angeles, CA • December 1987

DECENTRALIZED DYNAMIC DECISION MAKING †

H. R. Hashemi * and Ian B. Rhodes

Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106, USA

ABSTRACT

Problems of decision making in the face of statistical uncertainties have long been of interest to decision theorists in various disciplines. A decentralized detection or hypothesis-testing system is a special case of a decentralized decision-making system in which several local agents or detectors observe a common region of the environment to determine, for example, the presence or absence of a certain phenomenon. Due to such factors as communication constraints, local detectors must make decisions at their local sites and send their decisions, rather than their received measurements or observations, to a central processor, who is responsible for declaring the final decision. Moreover, the local detectors are not permitted to communicate with one another. In this paper, we study a broad class of decentralized multi-stage, multi-detector binary hypothesis-testing problems. It is shown that, under appropriate independence assumptions on the received measurements, local strategies at each time instant are given by threshold tests on the likelihood ratio. Furthermore, it is shown that local decisions of each detector depend not only on his present and past observations, but on his past decisions as well. That is, for each local detector there is a different threshold corresponding to each combination of past decisions.

I. INTRODUCTION

Problems of decision making in the face of statistical uncertainties have long been of interest to decision theorists in various disciplines. A great deal of attention has been given to these problems in order to cope with uncertainty. Detection theory is one area of decision making that has received much attention, particularly in surveillance systems. The well-known theory of classical detection (e.g., [Van Trees, 1968, 1971]) has been motivated by single-sensor detection problems. The situation is quite different when physically distributed sensors obtain measurements from a common region of the environment in order to determine the presence or absence of a certain phenomenon.

Even though classical detection theory is equally applicable to such multi-sensor detection problems, in practice there may be a need to communicate received raw data from geographically dispersed locations to a central processing unit. Such communication capability might not be feasible for a number of reasons that we will discuss shortly. In such a case, each of the local processors may make a local decision and send this decision, rather than the measured data, to the supervisor or fusion center where these local decisions are used to make an overall final decision. This problem is sometimes referred to as a decentralized (or distributed) detection problem since central processing of all measurements is not available.

Distributed detection is gaining increasing practical significance, especially in surveillance systems. Although the performance of decentralized detection is suboptimal compared to centralized detection, it is in many cases a more realistic formulation in practical applications than its centralized counterpart. For example, as discussed above, in the face of capacity-constrained channels, local processing could substantially decrease the communication bandwidth to transmit raw data to a central site, thereby speeding up the process and reducing the transmission cost. Also, decentralization may be the natural way to model the problem in situations where there are multiple detectors of various nature or located at dispersed geographical sites. Furthermore, decentralized detection may just be imposed in situations where, due to enormous amounts of available raw data, centralized processing of the information is not feasible. Other issues include potential system reliability and integrity in the face of failures.

The problem of decentralized decision making can also be viewed in the framework of decentralized optimal control theory [Sandell, Varaiya, Athans, & Safonov, 1978], [Tenney & Sandell, 1981a, 1981b, 1981c]. However, unlike most decentralized control problems, the decentralized hypothesis-testing problem can be tackled in a relatively straightforward manner leading in many cases to explicit decision rules. This inherent simplicity is due to the absence of any feedback in the system [Tenney & Sandell, 1981a], [Tenney, 1982], so that decisions made by one local processor do not alter the system dynamics and have no influence on the information available to other processors.

From another standpoint, decentralized hypothesis testing can be viewed in the framework of team theory [Radner,

† This work was supported by the Office of Naval Research under Grant N00014-79-C-0549.

* H. R. Hashemi, formerly Hamid R. Hashemipour, is now in the Ph.D. → M.D. Program at the University of Miami.

1962], [Ho & Chu, 1972], [Ho, 1980]; a "team" of local processors take observations regarding a certain phenomenon, each of which must make an independent decision based on his own information and pass this decision on to a central processor, who is also a member of the team.

There are a variety of ways in which one can pose a decentralized detection problem. Each formulation has elements that capture one or all of the following three features of decentralized detection and decision-making problems:

(a) **Hierarchical Structure:** Local observations are processed and the result sent to a fusion center where a final or "global" decision is made. There may or may not be a direct communication between the local observers.

(b) **Multi-Stage or Sequential Structure:** Additional measurements can be taken by the local observers. Each local observer might wait until he elects to stop taking measurements before sending information to the fusion center, or information might be sent after each observation and the fusion center given the flexibility to determine when to declare a decision.

(c) **Information Rate or Bandwidth Reduction:** The preprocessing performed by each local observer should result in a significant reduction in the amount of data sent to the fusion center. Each local processor may, for example, make a decision at his local site and send his decision to the fusion center instead of, say, his observation (as in centralized detection) or even some sufficient statistics.

The object of this paper is to analyze a broad class of multi-stage distributed detection problems that extends the results of [Hashemi & Rhodes, 1987]. We are interested in limited communications due to such factors as limited channel capacity, transmission cost, etc. In other words, we are dealing with communication-constrained problems. In Sections III and IV, we shall thoroughly investigate a multi-stage, multi-person decentralized binary hypothesis-testing problem in which each detector, after obtaining each of his observations, sends a binary decision (0 or 1) to the fusion center or the "supervisor," who is given the option of declaring a final decision (i.e., deciding which of two possible hypotheses, H_0 or H_1, is true) at either stage depending on the local decisions that have been received. There is a cost associated with each combination of the true state and the final decision. In addition, a cost is incurred for delaying the final decision until the next time instant.

It is shown that, under appropriate independence assumptions, the optimal local strategies are governed by threshold tests on the likelihood ratio (i.e., likelihood-ratio tests), where the thresholds are off-line computable. This is the major contribution of this paper. The importance of this result is apparent when one recalls that decentralized detection problems are in general NP-complete [Tsitsiklis & Athans, 1985]. In other words, these independence assumptions make tractable an otherwise intractable problem.

The notation "$f(x) \underset{u=1}{\overset{u=0}{\gtrless}} t$" is the standard decision theory notation meaning that $u=0$ when $f(x)>t$, $u=1$ when $f(x)<t$, and $u=0$ or 1 when $f(x)=t$. For notational convenience, let $E_{x|y}[f(x)] \triangleq E[f(x)|y]$ denote the conditional expectation of $f(x)$ given y. For convenience, we shall use the notation $p(x|y)$ for $p_{X|Y}(x|y)$; i.e., the arguments are used to denote different functions. It is also assumed, for convenience and to simplify the development, that all conditional densities exist. W(will assume that the set of states of nature contains two elements, called hypothesis H_0 and hypothesis H_1, and that H refers to H_0 or H_1. The symbol "J" is used to denote the cost function, and λ to denote the likelihood-ratio function, $\lambda(X) = \frac{p(H_0|X)}{p(H_1|X)}$. The test $\lambda(X) \underset{u=1}{\overset{u=0}{\gtrless}} t$ is called a likelihood-ratio test and the constant t is called a threshold.

Briefly, the rest of the paper is organized as follows. In Section II, we briefly survey the existing decentralized decision-making algorithms in the literature. Section III introduces the problem statement and furnishes some preliminary development. The solution to the general decentralized detection problem is provided in Section IV. Finally, Section V presents some concluding remarks.

II. PRELIMINARIES

Several classes of decentralized detection problems have been studied over the past few years. A one-stage decentralized hypothesis-testing problem has been examined in [Tenney & Sandell, 1981a], in which each of the two local detectors takes a single observation, makes a local decision as to which of the two hypotheses is true and sends this decision to a "fusion center" when an overall decision is made. There is no communication between the local detectors and their observations are statistically independent. The person-by-person optimal local strategies shown to be given by threshold tests on the likelihood ratio. Some extensions of the above problem have also been reported in [Lauer & Sandell, 1982a] and [Lauer & Sandell, 1982b]. In [Ekchian & Tenney, 1982] a one-step team decision problem has been considered where communications are allowed among the decision agents under a prearranged causal ordering. The decision rules are likelihood ratios based on the data, with thresholds determined by incoming communicated messages.

The above problems are restricted to only one time step. Several authors have considered decentralized multi-step problems. In [Kushner & Pacut, 1982], the information communicated by each of the two local detectors to the fusion center is a conditional probability; this serves as a sufficient statistic, so that from the coordinator's point of view the problem is a centralized one and his optimal strategy is a likelihood ratio test. Since transmitted conditional probabilities are real numbers, this scheme does not substantially reduce the communication bandwidth unless the number of observations made by each local detector is large.

In [Teneketzis, 1982], [Teneketzis & Ho, 1985] a decentralized extension of the classical Wald problem [Wald, 1947], has been studied and solved via dynamic programming. At every time instant, each of the two detectors decides whether to stop and transmit to the fusion center his decision as to which of the two hypotheses is true, or to continue until the next time step. A cost is incurred for taking each additional observation. The person-by-person optimal strategies are given by likelihood-ratio tests at each time instant. The thresholds of the two detectors are coupled and can be determined by the solution of a set of nonlinear algebraic equations.

In [Hashemi & Rhodes, 1987], a two-step, two-detector decentralized detection problem has been considered, in which a central processor, after receiving the two local decisions at time 1, decides whether to terminate the process and declare a decision or to continue to time 2, by which point he must declare a decision. Local strategies at both time instants are shown to be given by likelihood-ratio tests. The class of problems treated here and in [Hashemi & Rhodes, 1987] differs from that in [Teneketzis, 1982] or [Teneketzis & Ho, 1985], by giving to the fusion center the responsibility for determining when the final overall decision is declared. Unlike in [Kushner & Pacut, 1982], the local decisions are binary (0 or 1) and provide the supervisor with greatly reduced information about the local observations.

For several classes of distributed team decision problems, conditions for asymptotic convergence of each agent's decision sequence and asymptotic agreement (consensus) among all agents' decisions have been reported [Tsitsiklis & Athans, 1984], [Washburn & Teneketzis, 1984].

III. PROBLEM STATEMENT

3.1.Introduction

In this section, we pose a broad class of distributed detection problems that possess a sequential nature. Our formulation captures elements of all three features of distributed decision making problems that we listed in Section I. The hierarchical structure involves many local observers who communicate only with the single supervisor or "fusion center" and do not communicate among themselves. We consider a multi-stage problem in which the local decision makers communicate with the supervisor at each time step and it is the supervisor who determines when the process is terminated and a final overall decision is declared, c.f. [Teneketzis, 1982], [Teneketzis & Ho, 1985], To incorporate communication constraints we consider the extreme case where at each time each detector makes a binary decision (0 or 1) and sends only this (and not his observation) to the supervisor. It is assumed that all transmissions are error free. We adopt a Bayesian approach in which the common goal of the local decision makers and the supervisor is to minimize the expectation of a cost functional that reflects the correctness of the final decision and the time taken to reach it. Because direct communication between local decision makers is prohibited, the cooperation and indirect interaction among them is manifested only through this common team goal [Radner, 1962], [Ho, 1980].

An essential part of any decentralized decision or control problem is the specification of the information pattern [Witsenhausen, 1971], viz. the information that is available to each decision maker at each time instant. We consider that each local decision maker has access only to his past and present observations and his past decisions (equivalently, to his past and present observations and to his past decision rules, from which he can reconstruct his past decisions), i.e. each local decision maker, considered separately, has a *classical information pattern*. The supervisor has access only to the past and present local decisions communicated to him.

Without additional structure this decentralized problem is NP-complete. By assuming that the local observations are independent when conditioned on the hypotheses, the problem can be shown to be tractable. In other words, the problem is not NP-complete and can thus be solved in polynomial time [Hashemi & Rhodes, 1987], [Papadimitriou & Tsitsiklis, 1982], [Tsitsiklis & Athans, 1985], [Garey & Johnson, 1979]. This independence assumption is reasonable in many practical applications (e.g., in problems of detecting a known signal in uncorrelated noise) although it will not be valid in some other situations (e.g., in problems of detecting unknown signals). The additional structure added to the problem is therefore not too restrictive and permits explicit computations of the decision rules.

3.2. Problem Statement

Suppose there are two possible hypotheses, H_0 and H_1, with given *a priori* probabilities $p_j=P\{H_j\}$, j=0,1. There are K detectors (control stations), $U^1,\ldots,U^K$, each operating for N time steps. Detector U^k takes an observation y_n^k at time n and, based on his present and past observations $(y_1^k,\ldots,y_n^k)$ and his past decisions $(u_1^k,\ldots,u_{n-1}^k)$, sends a binary decision $u_n^k\in\{0,1\}$ to the supervisor, who is given the flexibility of either stopping and declaring a "global" decision or continuing to the next time step. (He is, however, forced to declare a decision at or prior to the final time.) This is an extreme case of limited communication between the local detectors and the supervisor. There is no communication among the local detectors.

To specify the information pattern, define the following sets for $1\le n\le N$ and $1\le k\le K$ (the following notation is adopted from [Witsenhausen, 1971]):

$$U_{n,k}=\{1,2,\ldots,n-1\}\ ,\quad U_{1,k}=\varnothing \tag{3.1a}$$

$$Y_{n,k}=\{1,2,\ldots,n\} \tag{3.1b}$$

so that

$$u_{U_{n,k}}\triangleq(u_1^k,\ldots,u_{n-1}^k)\ ,\quad y_{Y_{n,k}}\triangleq(y_1^k,\ldots,y_n^k) \tag{3.1c}$$

Also define

$$y^k\triangleq(y_1^k,\ldots,y_N^k)=y_{Y_{N,k}}\ ,\quad u^k\triangleq(u_1^k,\ldots,u_N^k)=u_{U_{N+1,k}} \tag{3.1d}$$

The objective is to find optimal local strategies for each detector (γ_n^k, with $1\le n\le N$ and $1\le k\le K$) and optimal strategies for the supervisor (γ_n, $1\le n\le N$) so as to minimize the cost J(Γ), where

$$\Gamma\triangleq(\gamma_n^k,\gamma_n:1\le n\le N,\ 1\le k\le K)\cdot \tag{3.2}$$

The local decisions are determined as

$$u_n^k=\gamma_n^k(y_{Y_{n,k}},u_{U_{n,k}})\ ,\quad k=1,\ldots,K \tag{3.3}$$

In other words, each detector has *perfect recall*. Fig. 3.1 shows the network topology for this problem and the information structure of the kth detector at time n. The global decision is given by

$$u_n=\gamma_n(u_{U_{n+1,k}}\,;k=1,\ldots,K) \tag{3.4}$$

It is assumed that the joint conditional p.d.f. $p(y^1,\ldots,y^K|H)$ is known *a priori*. There is a cost associated with each combination of the true state of nature and the global decision (corresponding, for example, to false alarm, missed detection, etc.). The cost incurred by declaring H_i (i.e., $u=i$) when H_j is true is denoted by $c_{ij}\ge0$. We assume that $c_{10}\ge c_{00}$ and $c_{01}\ge c_{11}$; that is, the cost of erring is at least as large as the cost of no error when the same hypothesis is true. In addition, there is an additive delay cost of $c_0\ge0$ if the supervisor postpones his decision until the next time instant.

3.3 Some Preliminary Results

First we introduce a crucial independence assumption which is used throughout the remainder of the paper.

Assumption 3.1

Provided the indicated p.d.f s exist, we assume that

$$p(y^1 y^2, \ldots, y^K | H) = p(y^1 | H)\, p(y^2 | H) \cdots p(y^K | H)$$

and

$$p(y^k | H) = p(y_1^k | H) \cdots p(y_N^k | H)$$

for both hypotheses $H{=}H_0$ and H_1 and all $1 \le k \le K$.

As was mentioned earlier, this assumption is reasonable in many practical situations (e.g., in problems of detecting a known signal in uncorrelated noise and in circumstances where the sensors are geographically dispersed). It results in the following lemma will be used in the next section.

Lemma 3.1

Let x and y be two random variables whose joint density function is given, and let $f(x,y)$ be a function of x and y. Suppose z is an extraneous random variable whose joint density function with x and y is well defined. Then

$$E\,[f(x,y)] = E_z\, E_{y|z}\, E_{x|y}\, [f(x,y)] \qquad (3.5)$$

Proof. It is obvious that, by using the smoothing property of conditional expectations, we have

$$E\,[f(x,y)] = E_{x,y}\,[f(x,y)] = E_{x,y,z}\,[f(x,y)] = E_z\, E_{y|z}\, E_{x|y,z}\,[f(x,y)]$$

We wish to show that the innermost expectation on the right side of this equation need not be conditioned on z. To see this, note that

$$E_{x,y}\,[f(x,y)] = E_y\, E_{x|y}\,[f(x,y)]$$

Now, let

$$g(y) = E_{x|y}\,[f(x,y)]$$

Then, after introducing the extraneous variable z and using the smoothing property, we obtain

$$E_{x,y}\,[f(x,y)] = E_y\, g(y) = E_z\, E_{y|z}\, g(y)$$

which is what we intended to show. □

One explanation for Lemma 2.2 is that f in the lemma is only a function of x and y, and not of z.

IV. RESULTS

In this section, we extend the results of [Hashemi & Rhodes, 1987] and solve the general multi-stage dynamic hypothesis-testing problem posed in the previous section. To find the local strategies, we pose the problem in the framework of a dynamic team problem and use the Bayes criterion to find the person-by-person optimal (PBPO) solutions [Ho & Chu, 1972], [Ho, 1980]. Our aim is to

$$\min_{\Gamma} \bar{J}(\Gamma) \triangleq \min_{\Gamma} E_{y^1,\ldots,y^K,H}\, J[\gamma(\gamma_n^k(y^k);\, n{=}1,\ldots,N;\, k{=}1,\ldots,K)] \qquad (4.1)$$

Eq. (4.1) is in the so-called *strategic form* of the problem. Using properties of nested conditional expectations and invoking Assumption 3.1, we obtain

$$\min_{\Gamma} E_{y^k}\, E_{H|y^k}\, E_{y^1,\ldots,y^{k-1},y^{k+1},\ldots,y^K|H}\, J_\gamma[\gamma^k(y^k);\, k{=}1,\ldots,K] \qquad (4.2)$$

where we have defined

$$J_\gamma[\gamma^k;\, k{=}1,\ldots,K] \triangleq J[\gamma(\gamma_n^k;\, n{=}1,\ldots,N;\, k{=}1,\ldots,K)] \qquad (4.3a)$$

and

$$\gamma^k \triangleq (\gamma_1^k, \ldots, \gamma_N^k) = \gamma_{Y_{n,k}} \qquad (4.3b)$$

Unfortunately, very little can be said about the globally optimal solution, and thus we focus our attention on PBPO solutions [Ho & Chu, 1972], [Ho, 1980]. To find, say, U^k's strategies, fix the strategies of all the other decision makers at their optima and minimize (4.2) over γ^k. Since the optimum strategies $\gamma^{*j}, j \ne k$, are given, the decisions $u^j, j \ne k$, become well-defined random variables and (4.2) reduces to the minimization over γ^k of

$$E_{y^k}\, E_{H|y^k}\, E_{u^1,\ldots,u^{k-1},u^{k+1},\ldots,u^K|H}\, J_\gamma[u^1,\ldots,u^{k-1},\gamma_n^k(y^k),u^{k+1},\ldots,u^K,H]$$

This is a non-trivial function minimization problem faced by U^k, but it can be reduced to the following equivalent parameter minimization problem by using the fact that finding the optimal u_n^k for every $y_{Y_{n,k}}$ and $u_{U_{n,k}}$ is equivalent to determining the optimal strategy γ^{*k}_n, with $1 \le n \le N$ [Ho, 1980]:

$$E_{y_1^k} \min_{u_1^k} E_{y_2^k|y_1^k} \min_{u_2^k|u_1^k} \cdots E_{y_N^k|y_{Y_{N-1,k}}} \min_{u_N^k u_{U_{N-1,k}}} E_{H|y^k}\, \Sigma_N^k(u^k,H) \qquad (4.4)$$

where we have defined

$$\Sigma_N^k(u^k,H) \triangleq E_{u^1,\ldots,u^{k-1},u^{k+1},\ldots,u^K|H}\, J_\gamma(u^1,\ldots,u^K,H) \qquad (4.5)$$

From (4.4) we can deduce U's strategy at the final time.

Strategies at time N

The inner minimization in (4.4) corresponds to the final time. Since $u_n^k \in \{0,1\}$, U^k's strategy at the final time is given by

$$E_{H|y^k}\, \Sigma_N^k(u_1^k,\ldots,u_{N-1}^k,1,H) \mathop{\gtrless}\limits_{u_N^k=1}^{u_N^k=0} E_{H|y^k}\, \Sigma_N^k(u_1^k,\ldots,u_{N-1}^k,0,H)$$

which, after expanding over H and rearranging, yields

$$p(H_0|y^k)\,[\Sigma_N^k(u_{U_{N,k}},1,H_0) - \Sigma_N^k(u_{U_{N,k}},0,H_0)]$$

$$\mathop{\gtrless}\limits_{u_N^k=1}^{u_N^k=0} p(H_0|y^k)\,[\Sigma_N^k(u_{U_{N,k}},0,H_1) - \Sigma_N^k(u_{U_{N,k}},1,H_1)] \qquad (4.6)$$

The above describes the PBPO strategy of Detector U^k at the final time. If we assume that the expression within the brackets on the left side of the above inequality is positive, then we can put (4.6) in a more convenient form, viz.

$$\lambda_N^k(y_{Y_{N,k}}) \triangleq \frac{p(H_0|y_{Y_{N,k}})}{p(H_1|y_{Y_{N,k}})} \mathop{\gtrless}\limits_{u_N^k=1}^{u_N^k=0} t_N^k(u_{U_{N,k}}) \qquad (4.7)$$

where

$$t_N^k(u_{U_{N,k}}) \triangleq \frac{\Sigma_N^k(u_{U_{N,k}},0,H_1) - \Sigma_N^k(u_{U_{N,k}},1,H_1)}{\Sigma_N^k(u_{U_{N,k}},1,H_0) - \Sigma_N^k(u_{U_{N,k}},0,H_0)} \qquad (4.8)$$

As can be seen, PBPO local strategies at time N are given by likelihood-ratio tests. We shall proceed to obtain the local strategies for time $n < N$.

Strategies at time $n < N$

We first prove the following lemma using the results of Lemma 3.1.

Lemma 4.1

Under Assumption 3.1, we hcve

$$E[J] = E_{u^1,\ldots,u^K,H} \{J_\gamma(u^1,\ldots,u^N,H)\} \qquad (4.9)$$

$$= E_{y_{Y_{n}}} E_{u_n^k|y_{Y_n}} E_{u_{U_{n-1}}|y_{Y_{n-1}}} E_{H|y_{Y_n}} E_{u_N^k,\ldots,u_{n+1}^k|u_{U_{n}},H} \Sigma_N^k(u^k,H)$$

That is, the innermost expectation need not be conditioned on $y_{Y_{n,k}}$.

Proof. After invoking Assumption 3.1, the proof parallels the proof of Lemma 3.1. □

It is important to note that since in (4.9) $u_{U_{n,k}}$ is availabie to U^k at time n (because U^k has perfect recall), we have

$$p(u_{U_{n,k}}|y_{Y_{n-1,k}}) = 1$$

for the available $u_{U_{n,k}}$. As a result, (4.9) further reduces to

$$E_{y_{Y_n}} E_{u_n^k|y_{Y_n}} E_{H|y_{Y_n}} \Sigma_n^k(u_{U_{n+1,k}},H) \qquad (4.10)$$

where we have defined

$$\Sigma_n^k(u_{U_{n+1,k}},H) \triangleq E_{u_N^k,\ldots,u_{n+1}^k|u_{U_{n}},H} \Sigma_N^k(u^k,H) \qquad (4.11)$$

Now, since $u_n^k \in \{0,1\}$ and $p(u_n^k{=}0|y_{Y_{n,k}}) = 1 - p(u_n^k{=}1|y_{Y_{n,k}})$, Eq. (4.10) can be written as

$$E_{y_{Y_n}} \sum_{u_n^k=0}^{1} \{p(u_n^k|y_{Y_{n,k}}) E_{H|y_{Y_n}} \Sigma_n^k(u_{U_{n+1,k}},H)\} \qquad (4.12)$$

$$= E_{y_{Y_n}} \left\{ p(u_n^k{=}1|y_{Y_{n,k}})\,\mu(y_{Y_{n,k}}) \right\} + E_H\ \Sigma_n^k(u_{U_{n,k}},0,H)$$

where we have defined

$$\mu(y_{Y_{n,k}}) = E_{H|y_{Y_n}} \{ \Sigma_n^k(u_{U_{n,k}},1,H) - \Sigma_n^k(u_{U_{n,k}},0,H) \}$$

Noticing that the last term in (4.12) is constant with respect to $p(u_n^k|y_{Y_{n,k}})$, the problem for Detector U^k at time n becomes one of minimizing $E_{y_{Y_n}} \{p(u_n^k{=}1|y_{Y_{n,k}})\,\mu(y_{Y_{n,k}})\}$. Clearly, this is minimized when

$$p(u_n^k{=}1|y_{Y_{n,k}}) = \begin{cases} 0 & \text{if } \mu(y_{Y_{n,k}}) > 0 \\ 1 & \text{if } \mu(y_{Y_{n,k}}) < 0 \end{cases}$$

or, equivalently, when

$$\mu(y_{Y_{n,k}}) \underset{u_n^k=1}{\overset{u_n^k=0}{\gtrless}} 0 \qquad (4.13)$$

After expanding $\mu(y_{Y_{n,k}})$ over H and collecting appropriate terms, we obtain

$$[\Sigma_n^k(u_{U_{n,k}},1,H_0)-\Sigma_n^k(u_{U_{n,k}},0,H_0)]\,p(H_0|y_{Y_{n,k}})$$

$$\underset{u_n^k=1}{\overset{u_n^k=1}{\gtrless}} [\Sigma_n^k(u_{U_{n,k}},0,H_1)-\Sigma_n^k(u_{U_{n,k}},1,H_1)]\,p(H_1|y_{Y_{n,k}}) \qquad (4.14)$$

The above describes the strategy of U^k at time n. As before, if we assume that the expression within the brackets on the left side of the above inequality is positive, then the PBPO strategy for U^k at time n can be represented in a more convenient way by a likelihood-ratio test, viz.

$$\lambda_n^k(y_{Y_{n,k}}) \triangleq \frac{p(H_0|y_{Y_{n,k}})}{p(H_1|y_{Y_{n,k}})} \underset{u_n^k=1}{\overset{u_n^k=0}{\gtrless}} t_n^k(u_{U_{n,k}}) \quad , \quad k=1,2,\ldots,K(4.15)$$

where the threshr ld is given by

$$t_n^k(u_{U_{n,k}}) \triangleq \frac{\Sigma_n^k(u_{U_{n,k}},0,H_1) - \Sigma_n^k(u_{U_{n,k}},1,H_1)}{\Sigma_n^k(u_{U_{n,k}},1,H_0) - \Sigma_n^k(u_{U_{n,k}},0,H_0)} \qquad (4.16)$$

(For $n{=}1$, $t_n^k(u_{U_{n,k}}){=}t_1^k$.)

In short, under Assumption 3.1, PBPO local strategies at each time instant have been shown to be described by threshold tests against the likelihood ratio.

Remark 4.1. It can be shown by natural extensions of [Hashemi & Rhodes, 1987] that each detector's thresholds at each time instant are coupled with the other detectors' thresholds at all time instants, as well as his own future thresholds; furthermore, they are parameterized by the supervisor's strategy. The coupling of the thresholds is indicated below. First define the threshold vector

$$\mathbf{t}^k \triangleq (t_n^k(u_{U_{n,k}}) : 1 \le n \le N) \quad , \quad k=1,\ldots,K \qquad (4.17)$$

Then, we have

$$t_n^k = \mathrm{f}_n^k(t_{n+1}^k,\ldots,t_N^k;\, \mathbf{t}^1,\ldots,\mathbf{t}^{k-1},\mathbf{t}^{k+1},\ldots,\mathbf{t}^K) \qquad (4.18)$$

or more compactly

$$\mathbf{t}^k = \mathbf{f}^k(\mathbf{t}^1,\ldots,\mathbf{t}^{k-1},\mathbf{t}^{k+1},\ldots,\mathbf{t}^K) \qquad (4.19)$$

In words, the thresholds are coupled with each other; however, thanks to Assumption 3.1, they can be computed off-line once and for all [by solving a system of nonlinear equations of the form (4.18)]. The functions f_n^k and $\mathbf{f}^k$ are also parametrized by the supervisor's strategy $\gamma = (\gamma_1, \ldots\ldots, \gamma_N)$.

Remark 4.2. It is interesting to see that each detector's thresholds at each time instant depend on his choice of past decisions; i.e., for each combination of past decisions, there is a different threshold for each detector. In other words, U^k's decision at time n depends not only on $y_{Y_{n,k}}$, but also on $u_{U_{n,k}}$. That is,

$$u_n^k = \gamma_n^k(y_{Y_{n,k}}, u_{U_{n,k}}) \qquad (4.20)$$

This leads to 2^{n-1} thresholds for U^k at time n, which are denoted $t_n^k(u_{U_{n,k}})$. Of course, once U^k's strategies at time n are given, then one can identify a region in the $y_1^k,\ldots,y_N^k$ hyperplane over which, say, $u_n^k{=}0$ is sent regardless of $u_{U_{n,k}}$. In Fig. 4.1, this situation is depicted at time 2. The decision regions at times 1 and 2 are shown for a typical example where a certain monotonicity is assumed (e.g., consider Gaussian distributions having the same standard deviations but different means under H_0 and H_1). It is important to emphasize here that these decision regions (i.e., the two regions $\{(y_1^k,y_2^k):u_2^k{=}i\}$, $i{=}0,1$) may not lend themselves to a threshold test. To put it another way, in general no single threshold can completely express U^k's optimal strategy a time 2. However, if one divides the y_1^k, y_2^k plane into two regions

$$A_{u_1^k} \triangleq \{(y_1^k,y_2^k) : \gamma_1^k(y_1^k) = u_1^k\} \qquad (4.21)$$

for $u_1^k{=}0,1$ as in Fig. 4.1, then over each of these two regions, A_0 and A_1, the decision u_2^k is decided via a likelihood-ratio test with thresholds $t_2^k(0)$ and $t_2^k(1)$, respectively. This is why at time 2 we need two thresholds and, in general, at time n we need 2^{n-1} thresholds for each detector.

Also note that the knowledge of y_1^k and γ_1^k is sufficient to reconstruct u_1^k. However, we have shown that U^k does not need to remember his past strategies, but rather his past decisions (which are easier to remember since they are just binary numbers). This simplification came about because we were able to reduce a function minimization problem to an equivalent parameter minimization problem (4.4).

Finally, observe that when N=2 in the above theorem, we get the results of [Hashemi & Rhodes, 1987], and when N=1, we obtain the strategies for the data fusion problem in [Tenney & Sandell, 1981a].

V. CONCLUSIONS

In this paper, we investigated a class of decentralized sequential decision-making problems. These problems dealt with the observing of a common region of the environment by a multiple of distributed sensors in order, for example, to detect the presence or absence of a certain phenomenon. We have extended the results of [Tenney & Sandell, 1981a], in which only one time step is considered, to a class of distributed detection problems that permit a sequential gathering of the observed data.

It was shown that, under the assumption that all measurements are statistically independent when conditioned on the hypotheses, each local strategy at each time instant is given by a threshold test on the likelihood ratio. Moreover, it was shown that the choice of a threshold by any detector at each step depends on the sequence of past decisions made by that detector.

Throughout this paper we have assumed that there are only two possible hypotheses and that each local processor can send one of two decisions (i.e., a binary message) to the central processor. It should be pointed out that this problem can be generalized to include multiple hypotheses (i.e., M-ary hypothesis testing) and multiple decisions (or multiple actions). The development is, of course, more tedious but straightforward.

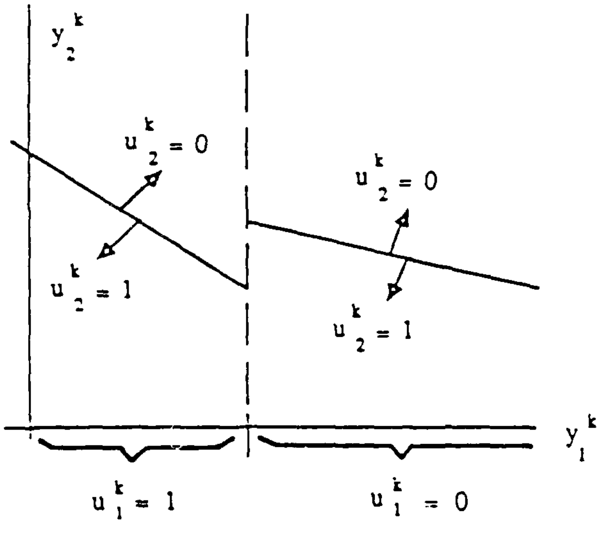


Fig. 4.1: The decision regions at time 1 and 2 for the kth local processor.

REFERENCES

Ekchian, L. K. and Tenney, R. R. (1982), "Detection Networks," *Proceedings of the 21st IEEE Conference on Decision and Control*, Orlando, Florida, December 8–10, 1982, pp. 686–691.

Hashemi, H. R. and Rhodes, I. B. (1987), "Decentralized Sequential Detection," submitted to the *IEEE Transactions on Information Theory*.

Ho, Y. C. (1980), "Team Decision Theory and Information Structures," *Proceedings of the IEEE*, Vol. 68, June 1980, pp. 644–654.

Ho, Y. C. and Chu, K. C. (1972), "Team Decision Theory and Information Structures in Optimal Control Problems-Part I," *IEEE Trarsactions on Automatic Control*, Vol. AC-17, No. 1, February 1972, pp. 15–22.

Kushner, H. J. and Pacut, A. (1982), "A Simulation Study of a Decentralized Detection Problem," *IEEE Transactions on Automatic Control*, Vol. AC-27, No. 5, October 1982, pp. 1116–1119.

Lauer, G. S. and Sandell, N. R., Jr. (1982a), "Distributed Detection of Signal Waveforms in Additive Gaussian Observation Noise," TP-160, ALPHATECH, Inc., Burlington, MA, 1982.

Lauer, G. S. and Sandell, N. R., Jr. (1982b), "Distributed Detection with Waveform Observations: Correlated Observation Processes," *Proceedings of the 1982 American Control Conference*, June 1982, pp. 812–819.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1982), "On the Complexity of Designing Distributed Protocols," *Information and Control*, Vol. 53, No. 3, June 1982, pp. 211–218.

Radner, R. (1962), "Team Decision Problems," *Annals of Mathematical Statistics*, Vol. 33, 1962, pp. 857–881.

Sandell, N. R., Jr., Varaiya, P., Athans, M., Safonov, M. G. (1978), "Survey of Decentralized Control Methods for Large-Scale Systems," *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, April 1978, pp. 108–128.

Teneketzis, D. (1982), "The Decentralized Wald Problem," *Proceedings of the 1982 IEEE International Large-Scale Systems Symposium*, Virginia Beach, VA, Oct 11–13, 1982, pp. 423–430.

Teneketzis, D. and Varaiya, P. (1984), "The Decentralized Quickest Detection Problem," *IEEE Transactions on Automatic Control*, Vol. AC-29, No. 7, July 1984, pp. 641–644.

Teneketzis, D. and Ho, Y. C. (1985), "The Decentralized Wald Problem," preprint.

Tenney, R. R. and Sandell, N. R., Jr. (1981a), "Detection with Distributed Sensors," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-17, July 1981, pp. 501–510.

Tenney, R. R. and Sandell, N. R., Jr. (1981b), "Structures for Distributed Decision Making," *IEEE Transacrions on Systems, Man, and Cybernetics*, Vol. SMC-11, No. 8, August 1981, pp. 517–527.

Tenney, R. R. and Sandell, N. R., Jr. (1981c), "Strategies for Distributed Decision Making," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-11, No. 8, August 1981, pp. 527–538.

Tsitsiklis, J. N. and Athans, M. (1984), "Convergence and Asymptotic Agreement in Distributed Decision Problems," *IEEE Transactions on Automatic Control*, Vol. AC-29, No. 1, January 1984, pp. 42–50.

Tsitsiklis, J. N. and Athans, M. (1985), "On the Complexity of Decentralized Decision Making and Detection Problems," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 5, May 1985, pp. 440–446.

Van Trees, H. L. (1968), *Detection, Estimation, and Modulation Theory–Part I*, Wiley, New York, 1968.

Van Trees, H. L. (1971), *Detection, Estimation, and Modulation Theory–Part III*, Wiley, New York, 1971.

Wald, A. (1947), *Sequential Analysis*, Wiley, New York, 1947.

Washburn, R. B. and Teneketzis, D. (1984), "Asymptotic Agreement Among Communicating Decision Makers," *Stochastics*, Vol. 13, 1984, pp. 103–129.

Witsenhausen, H. S. (1971), "Separation of Estimation and Control for Discrete Time Systems," *Proceedings of the IEEE*, Vol. 59, No. 11, November 1971, pp. 1557–1566.

Preprint of Paper:

"Decentralized Sequential Detection", H. R. Hashemi and Ian B. Rhodes, *IEEE Transactions on Information Theory,*, to appear.

Decentralized Sequential Detection *

HAMID R. HASHEMIPOUR† and IAN B. RHODES ‡

ABSTRACT

A class of decentralized sequential detection problems is investigated. Under appropriate independence assumptions, it is shown that at each time instant the optimal local strategies are given by threshold tests on the likelihood ratios. Furthermore, it is shown that local decisions depend not only on the present and past observations, but on the past local decisions as well. That is, for each local processor, there exists a different threshold for every different sequence of past decisions. Examples, computational techniques, discussions of the difficulties at hand, and suggestions for further exploration of the problem are presented.

* This work was supported by the Office of Naval Research under Grant N00014-79-C-0549.

† Formerly, with the Department of Electrical and Computer Engineering, University of California, Santa Barbara. He is now in the Ph.D. to M.D. program at the University of Miami School of Medicine.

‡ Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

I. Introduction

The problem of decentralized detection and decision making has been of increasing interest in recent years. This stems from the fact that, although it is suboptimal compared to centralized detection, decentralized detection is in many cases a more realistic formulation in practical applications than its centralized counterpart. To name a few, it could greatly decrease the communication bandwidth to transmit raw data to a central site, thereby speeding up the process and reducing the communication cost. Also, it may be the natural way to treat the problem when there are multiple detectors of various nature or located at different geographical sites. Furthermore, decentralized detection may just be imposed in situations where, due to enormous amounts of available raw data, centralized processing of the information is not feasible. Other issues include system reliability and integrity in the face of failures.

Several classes of decentralized detection problems have been studied over the past few years. In [1] a one-stage decentralized hypothesis-testing problem has been examined. Some variations of [1] have also been reported [2–4]. The multi-stage extension of the problem has also been studied for different classes of detection problems [5–8].

The object of this paper is the study of a more general and more complex class of multi-stage distributed detection problems. In Section II, we thoroughly investigate a two-step, two-person decentralized binary hypothesis-testing problem in which each detector, after obtaining each of its two observations, sends a binary decision (0 or 1) to the fusion center or the "supervisor," who is given the option of declaring the final decision (i.e., deciding which of two possible hypotheses, H_0 or H_1, is true) at either stage depending on the local decisions that have been received. It is shown that when all the observations are statistically independent given either hypothesis, the optimal local strategies are governed by Likelihood-Ratio Tests (LRT's) in a manner similar to centralized detection [9].

The class of problems treated in this paper is quite different from the one examined in [7–8], in which each detector is given the flexibility of either stopping and making a decision as to which hypothesis is true or continuing to the next time stage. In this paper, however, this flexibility is given to the supervisor; therefore, local detectors have no control over when to terminate their observation processes. Only after the supervisor has reached a decision can each local detector be notified to rest his process. The local decisions are binary decisions (0 or 1) which provide the supervisor with some information about their observations. (Unlike [5] this information is not served as sufficient statistics.)

This feature of the problem has an interesting implication. It turns out, as one might expect, that each detector's decision at each time instant depends not only on his past and present observations, but on his previous decisions (which he remembers) as well. For instance, at time two, there will be two thresholds for each detector corresponding to time-1 decisions of 0 and 1. Moreover, it is shown that each detector's thresholds at any time instant are coupled with the supervisor's strategy and the other detector's thresholds at all times, as well as his own future thresholds. They can be computed off-line and through the simultaneous solution of a system of nonlinear equations. Because local detectors are just sending binary decisions to the supervisor, it is not at all obvious, even under conditional independence assumptions, that the local strategies should be LRT's. The major contribution of this paper is to show that the local strategies are LRT's.

The generalization of the two-step problem to a multi-stage problem is straightforward and is discussed briefly at the end of Section II. Section III characterizes some of the important properties of the optimal solution. The results of this section suggest a significant reduction in the number of computations required to find the optimal solution. An example is presented in Section IV, and two techniques for solving it are discussed. The example illustrates the fact that the decision at any time instant for each detector depends on his past decisions (which are remembered by him as discussed above) as well as his present and past observations. Finally, Section V contains some concluding remarks and discusses some of the prospects for future research and the difficulties at hand.

Throughout this paper, we use the Bayesian criterion in determining the optimal rules. For notational convenience, let X^k denote the sequence of measurements, $X_1,\ldots,X_k$. Also, let $E_{X_2|X_1}[f(X_2)] \overset{\Delta}{=} E[f(X_2)|X_1]$ denote the conditional expectation of $f(X_2)$ given X_1. For convenience, we will use the notation $p(x,y|H)$ for $p_{X,Y|H}(x,y|H)$. That is, we use the arguments to denote different functions. We will assume that the set of states of nature contains two elements, called H_0 and H_1, and that H refers to H_0 or H_1. Unless otherwise indicated, i and j take the two values 0 and 1.

II. Two-Stage Decentralized Hypothesis Testing

There are a variety of ways in which one can pose a decentralized detection problem. A realistic formulation depends upon the concrete physical circumstances. In Problem 2.1 below, we investigate one class of two-stage decentralized detection problems as a vehicle for gaining insight into more general aspects of decentralized detection. Unlike [5], we are not interested in providing the coordinator with sufficient statistics; rather, due to such factors as communication constraints, each detector is asked to make a binary decision (0 or 1) at his local site and to send this binary message to the supervisor. Unlike [7–8], it is the supervisor here that decides when the process should be terminated; the local processors have no control over the stopping time. Naturally, we are interested in the manner by which detectors take the presence of one another and the supervisor into consideration. We shall use the Bayes criterion of optimality in our approach.

2.1. Problem Statement

There are two possible hypotheses, H_0 and H_1, with given a priori *probabilities $p_j=P\{H_j\}$, j=0,1, and two detectors U and V. There is no communication between the two detectors. Detector U (respectively, V) takes an observation X_1 (respectively, Y_1) at time 1 and makes a decision u_1 (respectively, v_1) based on his own observation, where u_1 and v_1 can take the two values 0 and 1. The decisions u_1 and v_1 are sent to the supervisor S, who either stops and makes a decision s_0 or waits until the second set of local decisions have arrived. In any case, U (respectively, V) obtains a second observation X_2 (respectively, Y_2) at time 2 and, based on his two measurements and his past decision, sends its decision u_2 (respectively, v_2) to the supervisor, where again u_2 and v_2 can take the two values 0 and 1. If the supervisor chose not to declare his decision s_0 at time 1, he would be obliged to do so at time 2 based on the four available local decisions $u^2 \overset{\Delta}{=} (u_1,u_2)$, and $v^2 \overset{\Delta}{=} (v_1,v_2)$.*

The joint pdf $p(X_1,X_2,Y_1,Y_2|H)$ is known a priori. The cost incurred by declaring H_i (i.e., s_0=i) when H_j is true is denoted by $c_{ij} \geq 0$. (We assume that $c_{10} \geq c_{00}$ and $c_{01} \geq c_{11}$; that is, the cost *of erring is at least as large as the cost of no error.) Besides, there is an additive delay cost of $c_0 \geq 0$ if the supervisor chooses to declare his decision at time 2.*[1] *It is assumed that at time n, X^n is available to U, Y^n to V, and $w^n \triangleq (u^n, v^n)$ to S, where $n \in \{1,2\}$. In addition, u_1 (respectively, v_1) is available to U (respectively, V) at time 2.*

Find optimal local strategies (ϕ_1 and ϕ_2 for detector U, ψ_1 and ψ_2 for detector V) and optimal rules for the supervisor (γ_1 and γ_2) so as to minimize the expected value of the overall system cost $J[\gamma(\phi,\psi),H]$ with respect to all present uncertainties (X^2, Y^2, and H) over all possible strategies ϕ, ψ, and γ, where $\gamma \triangleq (\gamma_1,\gamma_2)$, $\phi \triangleq (\phi_1,\phi_2)$, and $\psi \triangleq (\psi_1,\psi_2)$.

Solution

From the problem statement, it is understood that $u_1=\phi_1(X_1)$, $u_2=\phi_2(u_1,X^2)$, $v_1=\psi_1(Y_1)$, $v_2=\psi_2(v_1,Y^2)$, $s_1=\gamma_1(u_1,v_1)$, and $s_2=\gamma_2(u^2,v^2)$. So, for example, u_1 is measurable with respect to X_1, etc. Note that each agent, considered separately, has a classical information structure. In addition, since the detectors are not allowed to communicate, the information structure is static. The following independence assumption is used throughout this section:

Assumption 2.1

Observations X_1, X_2, Y_1, and Y_2 are mutually independent given H_0 or H_1. That is,

$$p(X^2,Y^2|H) = p(X^2|H)\, p(Y^2|H)$$

and

$$p(X^2|H) = p(X_1|H)\, p(X_2|H) \quad , \quad p(Y^2|H) = p(Y_1|H)\, p(Y_2|H) \; .$$

Let the S's decision at time n be denoted by s_n.[2] Fig. 2.1 depicts the decision tree for Problem 2.1 from the point of view of the supervisor S. We have $s_1 \in \{0,1,2\}$ and $s_2 \in \{0,1\}$, where $s_1=2$ implies that S waits for u_2 and v_2 to arrive. A moment of thought reveals that from the supervisor's viewpoint, the problem is totally centralized: he receives ''observations'' $w_1 \triangleq (u_1,v_1)$ at time 1 and, if he chooses to wait, $w_2 \triangleq (u_2,v_2)$ at time 2. Of course, S's ''observations'' depend on the decisions made by the decentralized decision-makers.

In this regard, once the local strategies are given, the decisions w_1 and w_2 become well-defined random variables whose joint conditional density function $p(w_1,w_2|H)$ can be determined. Consequently, the problem faced by S is centralized and can be solved using Fig. 2.1.

To find the local strategies, define $\Gamma \triangleq (\phi,\psi,\gamma)$ and pose the problem in the framework of a dynamic team problem [10]. Our aim is to

$$\min_{\Gamma} J(\Gamma) \triangleq \min_{\Gamma} E_{X^2,Y^2,H}\, J[\gamma(\phi(X^2),\psi(Y^2)),H] \tag{2.1a}$$

$$= \min_{\Gamma} E_{X^2}\, E_{H|X^2}\, E_{Y^2|H}\, J_\gamma[\phi(X^2),\psi(Y^2),H] \tag{2.1b}$$

[1] In other words, $J(s_0=i,H_j)=c_{ij}+c_0 \cdot I$ {decision reached at time 2}, where I is the indicator function.

[2] $s_0 \in \{0,1\}$ denotes the decision made by S once the process is *terminated*, where $s_0=i$ indicates that H_i has been declared. s_n refers to the decision made at time $n \in \{1,2\}$ with $s_1 \in \{0,1,2\}$ and $s_2 \in \{0,1\}$, where $s_1=2$ indicates that S continues to time 2.

where we have invoked Assumption 2.1 in the innermost expectation and have defined

$$J_\gamma(\phi,\psi,H) \triangleq J[\gamma(\phi,\psi),H)] \tag{2.1c}$$

There is very little known about the global optimal solution to (2.1). We thus focus our attention on person-by-person optimal (PBPO) solutions since a global solution, if one exists, must be PBPO. To this end, fix ψ^* and γ^* at their optima and minimize (2.1b) over ϕ. Since ψ^* is given, v_1 and v_2 are well-defined random variables, so

$$E_{Y^2|H}\ J_\gamma[\phi(X^2),\psi(Y^2),H] \equiv E_{v^2|H}\ J_\gamma[\phi(X^2),v^2,H] \tag{2.2}$$

Thus, (2.1b) becomes

$$\min_{\phi_1} \min_{\phi_2|\phi_1} E_{X^2}\, E_{H|X^2}\, E_{v^2|H}\ J_\gamma[\phi(X^2),v^2,H] \tag{2.3a}$$

$$= E_{X_1} \min_{u_1} E_{X_2|X_1} \min_{u_2|u_1} E_{H|X^2}\, E_{v^2|H}\ J_\gamma(u^2,v^2,H) \tag{2.3b}$$

where we have used the fact that finding the optimal u_2 for each u_1 and X^2 is equivalent to determining ϕ_2, and that finding the optimal u_1 for each X_1 is the same as determining ϕ_1 [11]. Eq. (2.3b) is in the so-called extensive form [10–11].

Strategies at time 2

The innermost minimization in (2.3b) corresponds to time 2. That is, given u_1 and X^2 at time 2, the problem facing detector U is to

$$\min_{u_2|u_1} E_{H|X^2}\, E_{v^2|H}\ J_\gamma(u^2,v^2,H) \triangleq \min_{u_2|u_1} E_{H|X^2}\, \Sigma_2(u^2,H) \tag{2.4a}$$

where we have defined

$$\Sigma_2(u^2,H) \triangleq E_{v^2|H}\ J_\gamma(u^2,v^2,H) \tag{2.4b}$$

Gathering appropriate terms in (2.4a) we get

$$p(H_0|X^2)\,[\Sigma_2(u_1,1,H_0)-\Sigma_2(u_1,0,H_0)] \underset{u_2=1}{\overset{u_2=0}{\gtrless}} p(H_1|X^2)\,[\Sigma_2(u_1,0,H_1)-\Sigma_2(u_1,1,H_1)] \tag{2.5}$$

where we use the standard decision theory notation

$$f(x) \underset{C_2}{\overset{C_1}{\gtrless}} t \iff \begin{cases} \text{Condition } C_1 \text{ is true if} & f(x)>t \\ \text{Condition } C_2 \text{ is true if} & f(x)<t \\ \text{Either condition is true if} & f(x)=t \end{cases}$$

Assume that

$$\Sigma_2(u_1,1,H_0) \geq \Sigma_2(u_1,0,H_0) \tag{2.6a}$$

$$\Sigma_2(u_1,0,H_1) \geq \Sigma_2(u_1,1,H_1) \tag{2.6b}$$

Then, the expressions within the square brackets on each side of (2.5) become nonnegative, and we can obtain the following strategy for detector U at time 2

$$\lambda_{u_2}(X^2) \underset{u_2=1}{\overset{u_2=0}{\gtrless}} t_{u_2}(u_1) \tag{2.7a}$$

where

$$t_{u_2}(u_1) \triangleq \frac{\Sigma_2(u_1,0,H_1)-\Sigma_2(u_1,1,H_1)}{\Sigma_2(u_1,1,H_0)-\Sigma_2(u_1,0,H_0)} \quad , \quad u_1=0,1 \tag{2.7b}$$

and the likelihood ratio λ_{u_2} is defined as

$$\lambda_{u_2}(X^2) \triangleq \frac{p(H_0|X^2)}{p(H_1|X^2)} \cdot \tag{2.7c}$$

In a similar manner, we get

$$\lambda_{v_2}(Y^2) \underset{v_2=1}{\overset{v_2=0}{\gtrless}} t_{v_2}(v_1) \tag{2.8a}$$

where

$$t_{v_2}(v_1) \triangleq \frac{\Xi_2(0,v_1,H_1)-\Xi_2(1,v_1,H_1)}{\Xi_2(1,v_1,H_0)-\Xi_2(0,v_1,H_0)} \quad , \quad v_1=0,1 \tag{2.8b}$$

and

$$\Xi_2(v^2,H) \triangleq \mathrm{E}_{u^2|H}\ \mathrm{J}_\gamma(u^2,v^2,H) \cdot \tag{2.8c}$$

Note that, due to Assumption 2.1, t_{u_2} (respectively, t_{v_2}) is independent of X^2 (respectively, Y^2). Consequently, the local strategies at time 2 are described by LRT's. Furthermore, it is interesting to see that t_{u_2} and t_{v_2} depend on the past decisions u_1 and v_1, respectively. That is, there are two thresholds for each detector at time 2, corresponding to time-1 decisions of 0 and 1, so that in general no single threshold can fully describe the strategy of a detector at time 2.

It can also be shown that U's strategy at time 2 depends on V's, as well as the supervisor's, strategies at time 1 and 2. This will be discussed later. We have thus determined the optimal local strategies ϕ_2 and ψ_2 at time 2 given (ϕ_1,ψ,γ) and (ϕ,ψ_1,γ), respectively.[3]

Strategies at time 1

To find optimal local strategies at time 1, we first prove the following lemma:

Lemma 2.1

Let x and y be two random variables whose joint density function is given, and let $f(x,y)$ be a function of x and y. Suppose z is an extraneous random variable whose joint density function with x and y is well defined. Then

$$\mathrm{E}\,[f(x,y)] = \mathrm{E}_z\ \mathrm{E}_{y|z}\ \mathrm{E}_{x|y}\ [f(x,y)] \tag{2.9}$$

Proof

It is obvious that, by using nested conditional expectations, we have

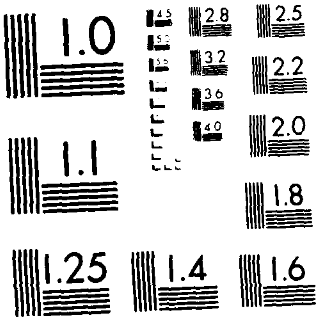
$$\mathrm{E}\,[f(x,y)] = \mathrm{E}_{x,y}\ [f(x,y)] = \mathrm{E}_{x,y,z}\ [f(x,y)] = \mathrm{E}_z\ \mathrm{E}_{y|z}\ \mathrm{E}_{x|y,z}\ [f(x,y)]$$

We wish to show that the innermost expectation need not be conditioned on z. To see [illegible] that

[3] In fact, as we shall discuss in Remark 2.5, ϕ_1 and ψ_1 can be replaced by u_1 and v_1, respectively.

AD-A194 163 PERFORMANCE OF STOCHASTIC AND DECENTRALIZED SYSTEMS(U) 2/2
CALIFORNIA UNIV SANTA BARBARA DEPT OF ELECTRICAL AND
COMPUTER ENGINEERING I B RHODES 11 FEB 88
UNCLASSIFIED N00014-59-C-0459 F/G 25/5 NL

END
DATE
FILMED
8 88



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

$$\mathrm{E}_{x,y}\ [f\,(x,y)] = \mathrm{E}_y\ \mathrm{E}_{x|y}\ [f\,(x,y)]$$

Now, let

$$g(y) = \mathrm{E}_{x|y}\ [f\,(x,y)]$$

Then, after introducing the extraneous variable z, we obtain

$$\mathrm{E}_{x,y}\ [f\,(x,y)] = \mathrm{E}_y\ g(y) = \mathrm{E}_z\ \mathrm{E}_{y|z}\ g(y)$$

which is what we intended to show. □

One explanation for Lemma 2.1 is that f in the lemma is only a function of x and y, and not of z. We can now prove the following corollary:

Corollary 2.1

Under Assumption 2.1, we have

$$\mathrm{E[J]} = \mathrm{E}[\mathrm{J}_\gamma\,(u^2,v^2,H)] \tag{2.10a}$$

$$= \mathrm{E}_{X_1}\ \mathrm{E}_{u_1|X_1}\ \mathrm{E}_{H|X_1}\ \mathrm{E}_{u_2|u_1,H}\ \Sigma_2(u^2,H) \tag{2.10b}$$

Proof

Again, here, the innermost expectation need not be conditioned on X_1. To show this, invoke Assumption 2.1 and properties of nested expectations to get

$$\mathrm{E[J]} = \mathrm{E}_{u^2,H}\ \mathrm{E}_{v^2|H}\ \mathrm{J}_\gamma(u^2,v^2,H) = \mathrm{E}_{u_1,H}\ \mathrm{E}_{u_2|u_1,H}\ \Sigma_2(u^2,H)$$

As in the lemma, define the function g to be

$$g(u_1,H) = \mathrm{E}_{u_2|u_1,H}\ \Sigma_2(u^2,H)$$

and introduce the "extraneous" random variable X_1 to get

$$E[\mathrm{J}] = \mathrm{E}_{u_1,H}\ g(u_1,H) = \mathrm{E}_{X_1}\ \mathrm{E}_{u_1,H|X_1}\ g(u_1,H) = \mathrm{E}_{X_1}\ \mathrm{E}_{u_1|X_1}\ \mathrm{E}_{H|X_1}\ g(u_1,H)$$

This completes the proof. □

Now, expand (2.10b) over u_1 to get

$$\mathrm{E}_{X_1} \sum_{u_1=0}^{1} p(u_1|X_1)\ \mathrm{E}_{H|X_1}\ \mathrm{E}_{u_2|u_1,H}\ \Sigma_2(u^2,H) \tag{2.11a}$$

$$= \mathrm{E}_{X_1}\, p(u_1{=}0|X_1)\ [\ \mathrm{E}_{H|X_1}\ \mathrm{E}_{u_2|u_1=0,H}\ \Sigma_2(0,u_2,H)\ ]$$

$$+ \mathrm{E}_{X_1}\, p(u_1{=}1|X_1)\ [\ \mathrm{E}_{H|X_1}\ \mathrm{E}_{u_2|u_1=1,H}\ \Sigma_2(1,u_2,H)\ ] \tag{2.11b}$$

$$= \mathrm{E}_{X_1}\left\{\ p(u_1{=}1|X_1)\left[\ \mathrm{E}_{H|X_1}\ \{\ \mathrm{E}_{u_2|u_1=1,H}\ \Sigma_2(1,u_2,H) - \mathrm{E}_{u_2|u_1=0,H}\ \Sigma_2(0,u_2,H)\ \}\ \right]\ \right\}$$

$$+ \mathrm{E}_H\ \mathrm{E}_{u_2|u_1=0,H}\ \Sigma_2(0,u_2,H) \tag{2.11c}$$

where we have used the fact that $p(u_1{=}0|X_1){=}1{-}p(u_1{=}1|X_1)$. Denote the expression within the square brackets in (2.11c) by $\mu(X_1)$. Since the last term in (2.11c) is constant with respect to $p(u_1|X_1)$, to minimize the total cost, we must minimize

$$\mathrm{E}_{X_1}\ \{\ p(u_1{=}1|X_1)\ \mu(X_1)\ \} \tag{2.11d}$$

Now, since $p(u_1{=}1|X_1) \in \{0,1\}$, it is obvious that to minimize the above we must have

$$p(u_1{=}1|X_1) = \begin{cases} 0 & \mu(X_1) > 0 \\ 1 & \mu(X_1) < 0 \end{cases} \tag{2.11e}$$

That is, we have

$$\mu(X_1) \underset{u_1=1}{\overset{u_1=0}{\gtrless}} 0 \tag{2.11f}$$

After expanding $\mu(X_1)$ over H and collecting appropriate terms, we obtain

$$p(H_0|X_1)\,[\Sigma_1(1,H_0){-}\Sigma_1(0,H_0)] \underset{u_1=1}{\overset{u_1=0}{\gtrless}} p(H_1|X_1)\,[\Sigma_1(0,H_1){-}\Sigma_1(1,H_1)] \tag{2.11g}$$

where we have defined

$$\Sigma_1(u_1,H) \triangleq \mathrm{E}_{u_2|u_1,H}\ \Sigma_2(u^2,H) \tag{2.11h}$$

Assume that

$$\Sigma_1(1,H_0) \ge \Sigma_1(0,H_0) \tag{2.12a}$$

$$\Sigma_1(0,H_1) \ge \Sigma_1(1,H_1) \tag{2.12b}$$

Then, the expressions within the square brackets on each side of (2.11g) become nonnegative, and we can obtain the following strategy for detector U at time 1

$$\lambda_{u_1}(X_1) \underset{u_1=1}{\overset{u_1=0}{\gtrless}} t_{u_1} \tag{2.13a}$$

where

$$t_{u_1} \triangleq \frac{\Sigma_1(0,H_1){-}\Sigma_1(1,H_1)}{\Sigma_1(1,H_0){-}\Sigma_1(0,H_0)} \tag{2.13b}$$

with the usual definition for the LR (Likelihood Ratio). Similarly, for V

$$\lambda_{v_1}(Y_1) \underset{v_1=1}{\overset{v_1=0}{\gtrless}} t_{v_1} \tag{2.14a}$$

where

$$t_{v_1} \triangleq \frac{\Xi_1(0,H_1){-}\Xi_1(1,H_1)}{\Xi_1(1,H_0){-}\Xi_1(0,H_0)} \tag{2.14b}$$

$$\Xi_1(v_1,H) \triangleq \mathrm{E}_{v_2|v_1,H}\ \Xi_2(v^2,H)\ \cdot \tag{2.14c}$$

It is crucial to note that, because of the presence of the term $\mathrm{E}_{u_2|u_1,H}$ in the definition of Σ_1 in (2.11h), t_{u_1} in (2.13) depends on ϕ_1. However, if ϕ_1 is fixed at its optimum $\phi_1{}^*$ (and all other strategies are fixed), then $t_{u_1}(\phi_1)$ becomes a constant, whereby (2.13a) lends itself to a LRT. The whole point is that t_{u_1} as defined by (2.13b) is functionally independent of X_1, so that an off-line calculation (as explained below in Remark 2.4) is possible.

Consequently, decision rules given by (2.13)–(2.14) above are LRT's, and they define the optimal strategies ϕ_1 and ψ_1 for time 1 (given that all the other strategies are fixed at their optima). We can now show that the right-hand side in (2.7b) depends on t_{v_1}, $t_{v_2}(0)$, and $t_{v_2}(1)$; i.e., U's thresholds at time 2 are coupled with V's thresholds at time 1 and 2. (A similar argument holds for V's thresholds at time 2.) To see this, use the definition of $t_{u_2}(u_1)$ in (2.7b) and note that in (2.4b) we have, for example,

$$p(v^2{=}00|H) = \iint_{\{Y^2:\ \lambda_{v_2}(Y^2)\geq t_{v_2}(0) \text{ and } \lambda_{v_1}(Y_1)\geq t_{v_1}\}} p(Y_2|H)\, p(Y_1|H)\, dY_1\, dY_2 \cdot$$

As a result, we obtain

$$t_{u_2}(u_1) = \mathrm{f}_{u_2}(t_{v_1}, t_{v_2}(0), t_{v_2}(1)) \quad , \quad u_1{=}0,1 \tag{2.15a}$$

$$t_{v_2}(v_1) = \mathrm{f}_{v_2}(t_{u_1}, t_{u_2}(0), t_{u_2}(1)) \quad , \quad v_1{=}0,1 \tag{2.15b}$$

It can also be shown that the right-hand side in (2.13b) depends on t_{v_1}, $t_{v_2}(0)$, $t_{v_2}(1)$, $t_{u_2}(0)$, $t_{u_2}(1)$, and t_{u_1}; i.e., U's threshold at time 1 is coupled with V's thresholds at time 1 and 2, as well as his own thresholds at time 2. (A similar argument holds for t_{v_1}, too.) To see this, use the definition of t_{u_1} in (2.13b) and Σ_1 in (2.11h). We then obtain

$$t_{u_1} = \mathrm{f}_{u_1}(t_{u_1}, t_{u_2}(0), t_{u_2}(1), t_{v_1}, t_{v_2}(0), t_{v_2}(1)) \tag{2.16a}$$

$$t_{v_1} = \mathrm{f}_{v_1}(t_{v_1}, t_{v_2}(0), t_{v_2}(1), t_{u_1}, t_{u_2}(0), t_{u_2}(1)) \,. \tag{2.16b}$$

Remarks

2.1. Note that f_{u_1} itself is a function of t_{u_1} because of the presence of the term $\mathrm{E}_{u_2|u_1,H}$ in the definition of Σ_1 in (2.11h). Similarly, f_{v_1} is a function of t_{v_1}. So, for example, if $t_{u_2}(0)$, $t_{u_2}(1)$, t_{v_1}, $t_{v_2}(0)$, and $t_{v_2}(1)$ are fixed at their optima, then the optimum t_{u_1} can be determined by finding the fixed point of f_{u_1} in (2.16a).

2.2. It should be pointed out that the functions in (2.15) and (2.16) are parameterized by γ. However, once γ is fixed at its optimum, local thresholds can be determined through the simultaneous solution of six nonlinear algebraic equations of the form (2.15)–(2.16) in six unknowns. Since the total possible choices of γ is finite, one can find the optimal thresholds through (2.15)–(2.16) for each choice of γ and pick the one(s) that correspond to the smallest cost. Each solution will result in a PBPO solution to Problem 2.1.

2.3. We have shown that the local strategies for any globally optimum or PBPO solution must be LRT's against a set of precomputable thresholds. Although little can be said about the existence of either global or PBPO solutions, we can show that a globally optimum (and therefore a person-by-person optimal) solution exists if the local strategies are restricted to being LRT's against thresholds. This globally optimum solution under this constraint must either be globally optimum in the absence of constraints or no global optimum exists.

To show the existence of a globally optimum solution when local strategies are restricted to LRT's against thresholds, note that a LRT of the form

$$\frac{p(H_0|X)}{p(H_1|X)} \underset{u=1}{\overset{u=0}{\gtrless}} t$$

is equivalent to the following LT (Likelihood Test)

$$p(H_0|X) \underset{u=1}{\overset{u=0}{\gtrless}} \tau$$

where $\tau{=}t/(1{+}t) \in [0,1]$. We refer to a solution in which local strategies are restricted to LT's against thresholds as a "threshold" solution. Now, fix the supervisor's strategy γ satisfying assumptions (2.6)–(2.7), and similar ones for the other detector. (These assumptions are equivalent to Corollary 3.--- below.) Since for such a γ PBPO solutions are LT's, we can focus our attention on the existence of optimal thresholds, τ_{u_1}, etc. Define

$$L(\tau) = \mathrm{E}\{\mathrm{J}_\gamma(u^2, v^2, H)\}$$

where $\tau \triangleq (\tau_{u_1}, \tau_{u_2}(0), \tau_{u_2}(1), \tau_{v_1}, \tau_{v_2}(0), \tau_{v_2}(1))$. Since each τ lies in the interval [0,1], the domain of L is a unit hypercube (which is a compact set). The cost function L can be written as a sum of 32 terms corresponding to 2^5 possible sequences of (u^2, v^2, H). Each of these terms involves integrals over regions that are functions of the thresholds, τ.

Assuming that the *a priori* pdf's are piecewise continuous in the observations, it is clear that these integrals are jointly continuous in the thresholds. Consequently, L is jointly continuous in the six thresholds. That is, L is a continuous function over a compact set. Therefore, by the Weierstrauss theorem L is bounded and assumes its minimum in that set. The minimum point then constitutes a "threshold" solution. It is clear that because of the finiteness of the supervisor's possible strategies, repeating this process to exhaust all enumerations of such γ's yields the globally optimal *threshold* solution. This "threshold" solution is globally optimal if a global optimum exists. If a globally optimal policy does not exist but a PBPO one does, then this threshold solution is a PBPO solution.

The preceding discussions can be summarized by the following theorem.

Theorem 2.1

Under Assumption 2.1, the person-by-person optimal solutions to Problem 2.1 are determined as follows:

(a) *Optimal local strategies ϕ_1 and ψ_1 at time 1 are described by LRT's:*

$$\lambda_{u_1}(X_1) \underset{u_1=1}{\overset{u_2=0}{\gtrless}} t_{u_1} \tag{2.17a}$$

$$\lambda_{v_1}(Y_1) \underset{v_1=1}{\overset{v_2=0}{\gtrless}} t_{v_1} \tag{2.17b}$$

The thresholds t_{u_1} and t_{v_1} must satisfy Eqs. (2.13b) and (2.14b).

(b) *Optimal local strategies ϕ_2 and ψ_2 at time 2 are also described by LRT's:*

$$\lambda_{u_2}(X^2) \underset{u_2=1}{\overset{u_2=0}{\gtrless}} t_{u_2}(u_1) \quad , \qquad u_1=0,1 \tag{2.18a}$$

$$\lambda_{v_2}(Y^2) \underset{v_2=1}{\overset{v_2=0}{\gtrless}} t_{v_2}(v_1) \quad , \qquad v_1=0,1 \tag{2.18b}$$

The thresholds $t_{u_2}(u_1)$ and $t_{v_2}(v_1)$ must satisfy Eqs. (2.7b) and (2.8b).

(c) *Supervisor's optimal strategy is based on a table look-up:*

$$s_1 = \gamma_1(u_1, v_1) \text{ and } s_2 = \gamma_2(u^2, v^2) \tag{2.19}$$

where γ_1 and γ_2 can be determined from Fig. 2.1 if the local thresholds are given.

Remarks

2.4. Due to Assumption 2.1, all thresholds in (a) and (b) can be computed off-line. As was discussed earlier, each detector's thresholds at time 2 are coupled with the other detector's thresholds at both times, and each detector's threshold at time 1 is coupled with the other detector's thresholds at both times, as well as his own thresholds at time 2; furthermore, these thresholds are coupled with the supervisor's optimal strategies γ. If we define the threshold vectors

$$(i)\ \mathbf{t}_u \triangleq (t_{u_1}, t_{u_2}(0), t_{u_2}(1)) \qquad (ii)\ \mathbf{t}_v \triangleq (t_{v_1}, t_{v_2}(0), t_{v_2}(1)) \tag{2.20}$$

then we have

$$(i)\ \mathbf{t}_u = \mathbf{f}_u^\gamma(\mathbf{t}_v) \qquad (ii)\ \mathbf{t}_v = \mathbf{f}_v^\gamma(\mathbf{t}_u) \tag{2.21}$$

where

$$\mathbf{f}_u^\gamma, \mathbf{f}_v^\gamma : \mathbf{R}^3 \to \mathbf{R}^3 . \tag{2.22}$$

In other words, U's and V's threshold vectors are coupled and can be found by solving two nonlinear algebraic vector equations of the form (2.21) in two unknowns, $\mathbf{t}_u$ and $\mathbf{t}_v$.

2.5. We have shown that, e.g., U's decision at time 2 depends not only on both his observations, X_1 and X_2, but also on his past decision, u_1. That is,

$$u_2 = \phi_2(u_1, X^2)$$

This resulted in two thresholds for U at time 2, $t_{u_2}(0)$ and $t_{u_2}(1)$. Of course, once U's strategies at time 2 are given, then one can identify a region in the X_1, X_2 plane over which, say, u_2=0 is sent regardless of u_1 (Fig. 2.2). It is important to point out that this decision region may not lend itself to a threshold test. To put it another way, in general no single threshold can completely express U's optimal strategy at time 2. However, if one divides the X_1, X_2 plane into two regions

$$A_{u_1} \triangleq \{(X_1, X_2) : \phi_1(X_1) = u_1\}$$

for u_1=0,1, then over each of these two regions, A_0 and A_1, the decision u_2 is decided via a LRT with thresholds $t_{u_2}(0)$ and $t_{u_2}(1)$, respectively. This is why two thresholds are needed for each detector at time 2.

2.2. Generalization

In this subsection, we state without proof the generalization of Problem 2.1 to a multi-stage hypothesis-testing problem. The problem can also be easily extended to include multiple detectors, hypotheses, and/or actions. As before, we make the following independence assumption:

Assumption 2.2

Suppose there are two detectors each taking N observations

$$X^N = (X_1, X_2, \ldots X_N) \quad and \quad Y^N = (Y_1, Y_2, \ldots, Y_N) .$$

We assume that

$$p(X^N|H) = p(X_1|H)\, p(X_2|H) \ldots p(X_N|H) ,$$

$$p(Y^N|H) = p(Y_1|H)\, p(Y_2|H) \ldots p(Y_N|H) ,$$

and

$$p(X^N, Y^N|H) = p(X^N|H)\, p(Y^N|H)$$

for all the hypotheses H in the set of states of nature.

The multiple-stage problem is identical to Problem 2.1, except we allow each detector to take N observations and send a decision after each observation is made. The cost incurred by declaring H_i when H_j is true is denoted by c_{ij}. There is also a non-negative additive delay cost of c_n ($1 \leq n \leq N$) when S waits for decisions u^n and v^n to arrive.[4] (In Problem 2.1 we had $c_1=0$ and $c_2=c_0$.) Again, each detector has access to all his past (and present) observations and past decisions.

Theorem 2.2

Under Assumption 2.2, the person-by-person optimal strategy of each local detector at every time step for the above problem is described by a LRT against a threshold, where the threshold depends on the choice of prior decisions of that local detector (but not on those of other local detectors). For instance, the optimal strategy of Detector U at time n is described by the following LRT's:

$$\lambda_{u_N}(X^N) \underset{u_N=1}{\overset{u_N=0}{\gtrless}} t_{u_N}(u^{N-1}) \tag{2.23a}$$

$$\lambda_{u_n}(X^n) \underset{u_n=1}{\overset{u_n=0}{\gtrless}} t_{u_n}(u^{n-1}) \ , \quad 1<n<N \tag{2.23b}$$

$$\lambda_{u_1}(X_1) \underset{u_1=1}{\overset{u_1=0}{\gtrless}} t_{u_1} \tag{2.23c}$$

where

$$t_{u_n}(u^{n-1}) \triangleq \frac{\Sigma_n(u^{n-1},0,H_1)-\Sigma_n(u^{n-1},1,H_1)}{\Sigma_n(u^{n-1},1,H_0)-\Sigma_n(u^{n-1},0,H_0)} \ , \quad 1 \leq n \leq N \tag{2.24a}$$

$$\Sigma_n(u^n,H) \triangleq \mathrm{E}_{u_N,\ldots,u_{n+1}|u^n,H} \ \Sigma_N(u^N,H) \ , \tag{2.24b}$$

and

$$\Sigma_N(u^N,H) \triangleq \mathrm{E}_{v^N|H} \ J_\gamma(u^N,v^N,H) \cdot \tag{2.24c}$$

Note that for every detector at time n there are 2^{n-1} thresholds, each corresponding to a different set of past decisions (which are remembered by that detector).

Again, it can be shown that each detector's thresholds at any time instant are coupled with the other detector's thresholds at all times and the supervisor's strategy, as well as his own future thresholds.

Finally, observe that when $N=1$ in the above theorem, we get the strategies for the data fusion problem in [1].

[4] We are supposing for completeness that S is allowed to reach a decision at time 0 based solely on the *a priori* probabilities, p_0 and p_1.

III. Characterization of the properties of optimal solutions

In this section, some useful properties associated with PBPO solutions as given by Theorem 2.1 are derived. Later on, we will see how these properties suggest a substantial reduction in the number of computations required for determining the optimal strategies.

Each PBPO solution is characterized by the properties that will shortly follow. The major results in this section are Theorems 3.1–3.3, which suggest a significant reduction in the number of possible enumerations of the optimal supervisor's strategy γ, thus making the computation of the solutions more feasible. We assume that thresholds lie in the extended positive real line $[0,\infty]$. Consequently, a threshold at ∞ implies that only one of the decisions, namely 1, gets sent irrespective of the actual observations. We start out by proving some preliminary lemmas. In the remainder of this section, the strategies are assumed to be PBPO as given by Theorem 2.1 (i.e., strategies are LRT's and assumed to be "monotonic" in that local decisions are identified with the hypotheses). For convenience, it is also assumed that $p(u_1,u_2|H)$, $p(v_1,v_2|H)$, and $p(H)$ are positive. This latter assumption simplifies the development that will follow shortly, but the results of Theorems 3.1–3.3 below still hold in the absence of this assumption.

Lemma 3.1

Let $\bar{t}_{u_1} \triangleq (p_1/p_0)t_{u_1}$. Then, for Detector U's PBPO strategy at time 1 we have

$$\frac{p(u_1{=}0|H_0)}{p(u_1{=}0|H_1)} \geq \bar{t}_{u_1} \geq \frac{p(u_1{=}1|H_0)}{p(u_1{=}1|H_1)}\,. \tag{3.1}$$

Proof

First note that

$$\frac{p(H_0|X_1)}{p(H_1|X_1)} \triangleq \lambda(X_1) \gtrless t \quad \Leftrightarrow \quad \frac{p(X_1|H_0)}{p(X_1|H_1)} \gtrless (p_1/p_0)\, t \triangleq \bar{t} \tag{3.2}$$

Using the PBPO local strategy for U given by Eq. (2.13), we have

$$p(u_1{=}0|H_0) = \int_{\{X_1:\lambda_{u_1}(X_1)\geq t_{u_1}\}} p(X_1|H_0)\, dX_1 \tag{3.3}$$

$$\geq \bar{t}_{u_1} \int_{\{X_1:\lambda_{u_1}(X_1)\geq t_{u_1}\}} p(X_1|H_1)\, dX_1 = \bar{t}_{u_1}\; p(u_1{=}0|H_1) \tag{3.4}$$

while

$$p(u_1{=}1|H_0) = \int_{\{X_1:\lambda_{u_1}(X_1)\leq t_{u_1}\}} p(X_1|H_0)\, dX_1 \tag{3.5}$$

$$\leq \bar{t}_{u_1} \int_{\{X_1:\lambda_{u_1}(X_1)\leq t_{u_1}\}} p(X_1|H_1)\, dX_1 = \bar{t}_{u_1}\; p(u_1{=}1|H_1) \tag{3.6}$$

The result is then immediate. □

One way of interpreting (3.1) is that the product of the conditional probabilities of two true decisions $p(u_1{=}0|H_0)\,p(u_1{=}1|H_1)$ is at least as large as the product of the conditional probabilities of two wrong decisions $p(u_1{=}0|H_1)\,p(u_1{=}1|H_0)$. The following corollary follows immediately from Lemma 3.1 and will be useful later on.

Corollary 3.1

For U's PBPO decision at time 1, we have

$$p(u_1{=}0 \mid H_0) \geq p(u_1{=}0) \geq p(u_1{=}0 \mid H_1) \tag{3.7a}$$

$$p(u_1{=}1 \mid H_1) \geq p(u_1{=}1) \geq p(u_1{=}1 \mid H_0) \tag{3.7b}$$

Clearly, there are counterparts to Lemma 3.1 and Corollary 3.1 for V. Corollary 3.1 has a very intuitive appeal: it says, in effect, that the probability of making an error, say, $p(u_1{=}0 \mid H_1)$, is no larger than the probability of declaring a true decision, say, $p(u_1{=}0 \mid H_0)$. Note that in the previous lemma, we have used the fact that local strategies are given by LRT's. Now we can prove the following lemma.

Lemma 3.2

For PBPO decisions u_1 and v_1 at time 1, we have

$$(i)\;\; p(H_0 \mid 0,v_1) \geq p(H_0 \mid 1,v_1) \qquad (ii)\;\; p(H_0 \mid u_1,0) \geq p(H_0 \mid u_1,1)$$

$$(iii)\;\; p(H_1 \mid 1,v_1) \geq p(H_1 \mid 0,v_1) \qquad (iv)\;\; p(H_1 \mid u_1,1) \geq p(H_1 \mid u_1,0)$$

Proof

To prove, say, inequality (i), invoke the Bayes' rule and use the fact that u_1 and v_1 are conditionally independent. Then, it must be shown that

$$\left[1+\frac{p(u_1{=}1 \mid H_1)p(v_1 \mid H_1)p1}{p(u_1{=}1 \mid H_0)p(v_1 \mid H_0)p0}\right]^{-1} \leq \left[1+\frac{p(u_1{=}0 \mid H_1)p(v_1 \mid H_1)p1}{p(u_1{=}0 \mid H_0)p(v_1 \mid H_0)p0}\right]^{-1} \tag{3.8}$$

and, after some simplifications, we need to show that

$$\frac{p(u_1{=}0 \mid H_0)}{p(u_1{=}0 \mid H_1)} \geq \frac{p(u_1{=}1 \mid H_0)}{p(u_1{=}1 \mid H_1)}\,. \tag{3.9}$$

But this is precisely Lemma 3.1. □

We are now in a position to prove the following important theorem. It will be discussed later how Theorem 3.1 and its companion Theorem 3.2 enhance our algorithm to compute the optimal thresholds. The set **C** is the set of all pairs (u_1,v_1) for which S continues to time 2.

Theorem 3.1

Let γ_1 define a person-by-person optimal strategy for the supervisor at time 1 as given by Theorem 2.1, and take $i,j,k,l \in \{0,1\}$ with $i \neq j$. We have

(a) *if $\gamma_1(i,j){=}j$, then $\gamma_1(j,j){=}j$ provided $(j,j) \notin \mathbf{C}$.*

(b) *if $\gamma_1(i,i){=}j$, then $\gamma_1(k,l){=}j$ provided $(k,l) \notin \mathbf{C}$.*

Proof

We prove (a) for $i=0$ and $j=1$. The generalization to other contingencies is straightforward. So, suppose S receives $(u_1,v_1)=(0,1)$ and declares $s_0=1$ at time 1. We will show that if $(u_1,v_1)=(1,1)$ is received by S, then, if he decides to stop at time 1, he must declare $s_0=1$ again. Since $\gamma_1(u_1v_1=01)=1$, using Fig. 2.2 we obtain

$$c_{10}\,p(H_0|u_1v_1=01) + c_{11}\,p(H_1|u_1v_1=01)$$

$$\le c_{00}\,p(H_0|u_1v_1=01) + c_{01}\,p(H_1|u_1v_1=01) \tag{3.10}$$

or, equivalently,

$$(c_{10}-c_{00})\,p(H_0|u_1v_1=01) \le (c_{01}-c_{11})\,p(H_1|u_1v_1=01) \tag{3.11}$$

where we have used the fact that $c_{10}\ge c_{11}$ and $c_{01}\ge c_{11}$. We wish to show that the above inequality also holds true for $(u_1,v_1)=11$. It suffices to show that

$$p(H_0|u_1v_1=11) \le p(H_0|u_1v_1=01) \tag{3.12a}$$

$$p(H_1|u_1v_1=01) \le p(H_1|u_1v_1=11) \tag{3.12b}$$

But (3.12) is precisely Lemma 3.2. This completes the proof. □

In simple terms, Theorem 3.1 indicates that if the supervisor stops and declares, say, a 0 upon receiving conflicting decisions (0,1 or 1,0) at time 1, then he must also stop and declare a 0 if he receives the nonconflicting decisions 0,0. This result is in part due to our "monotonicity" assumption stated in the beginning of the section.

Properties of Optimal Solutions at time 2

We shall now move on to time 2 and establish similar properties. The following lemma is the analogue of Lemma 3.1 for time 2, and its proof is similar to the proof of Lemma 3.1.

Lemma 3.3

Let $\bar{t}_{u_2}(u_1) \triangleq (p_1/p_0)t_{u_2}(u_1)$. Then, for Detector U's PBPO strategy at time 2 we have

$$\frac{p(u_1,u_2=0|H_0)}{p(u_1,u_2=0|H_1)} \ge \bar{t}_{u_2}(u_1) \ge \frac{p(u_1,u_2=1|H_0)}{p(u_1,u_2=1|H_1)}\,. \tag{3.13}$$

where $u_1=0,1$.

Lemma 3.3 can be interpreted in the same way as Lemma 3.1. We shall present without proof the following corollaries which will be useful later on. The proofs follow by applying Bayes' rule.

Corollary 3.2

Under the conditions of Lemma 3.3, we have

$$\frac{p(u_2=0|u_1,H_0)}{p(u_2=0|u_1,H_1)} \ge \bar{t}_{u_2}(u_1)\,\frac{p(u_1|H_1)}{p(u_1|H_0)} \ge \frac{p(u_2=1|u_1,H_0)}{p(u_2=1|u_1,H_1)} \tag{3.14}$$

Corollary 3.3

Under the conditions of Lemma 3.3, given u_1 at time 2, we have for u_2 that

$$p(u_2{=}0|u_1,H_0) \geq p(u_2{=}0|u_1) \geq p(u_2{=}0|u_1,H_1) \tag{3.15a}$$

$$p(u_2{=}1|u_1,H_1) \geq p(u_2{=}1|u_1) \geq p(u_2{=}1|u_1,H_0) \tag{3.15b}$$

Clearly, there are counterparts to Lemma 3.3 and its corollaries for V. Corollary 3.3 can be interpreted in the same way as Corollary 3.1. Combining Corollaries 3.1 and 3.3, we can get the following corollary.

Corollary 3.4

For Detector U's PBPO decisions at times 1 and 2, we have

$$p(u_1u_2{=}00|H_0) \geq p(u_1u_2{=}00) \geq p(u_1u_2{=}00|H_1) \tag{3.16a}$$

$$p(u_1u_2{=}11|H_1) \geq p(u_1u_2{=}11) \geq p(u_1u_2{=}11|H_0) \tag{3.16b}$$

We can now state the following lemma, whose proof is similar to the proof of Lemma 3.2.

Lemma 3.4

For PBPO decisions u^2,v^2 we have

(*i*) $p(H_0|u_1,0,v^2) \geq p(H_0|u_1,1,v^2)$ (*ii*) $p(H_0|u^2,v_1,0) \geq p(H_0|u^2,v_1,1)$

(*iii*) $p(H_1|u_1,1,v^2) \geq p(H_1|u_1,0,v^2)$ (*iv*) $p(H_1|u^2,v_1,1,) \geq p(H_1|u^2,v_1,0)$

In particular, for every pair (u_1,v_1), we have

(*v*) $p(H_0|u_1,v_1,0,0) \geq p(H_0|u_1,v_1,i,j) \geq p(H_0|u_1,v_1,1,1)$

(*vi*) $p(H_1|u_1,v_1,1,1) \geq p(H_1|u_1,v_1,i,j) \geq p(H_0|u_1,v_1,0,0)$

where $i,j \in \{0,1\}$ with $i \neq j$.

We are now in a position to establish the following important theorem, which is the analogue of Theorem 3.1 for time 2.

Theorem 3.2

The results of Theorem 3.1 are also applicable to $\gamma_2(u^2,v^2)$, the supervisor's PBPO strategy at time 2. Specifically, for any $(u_1,v_1) \in \mathbf{C}$ and $i,j,k,l \in \{0,1\}$ with $i \neq j$, we have

(a) *if $\gamma_2(u_1,v_1,i,j){=}j$, then $\gamma_2(u_1,v_1,j,j){=}j$.*

(b) *if $\gamma_2(u_1,v_1,i,i){=}j$, then $\gamma_2(u_1,v_1,k,l){=}j$.*

Proof

Use a line of argument similar to that in the proof of Theorem 3.1. □

In simple terms, Theorem 3.2 indicates that if the supervisor declares, say, a 0 upon receiving conflicting decisions (0,1 or 1,0) at time 2, then he must also declare a 0 if he receives the nonconflicting decisions 0,0. We can now prove the following theorem, which is another version of Theorem 3.2 but sheds some more light on the structure of the supervisor's optimal strategy.

Theorem 3.3

Theorem 3.2 can be stated differently as

$$J_\gamma(u_1,v_1,1,v_2,H_0) \geq J_\gamma(u_1,v_1,0,v_2,H_0) \qquad (3.17)$$

with similar inequalities for v_2 and for H_1.

Proof

There are two possible cases:

(A) The decision pair $(u_1,v_1) \in$ C: if $\gamma_2(u_1,v_1,0,v_2)=0$, then the RHS of (3.17) is c_0+c_{00} and, therefore, less than or equal to the LHS. But if $\gamma_2(u_1,v_1,0,v_2)=1$, then the RHS of (3.17) is c_0+c_{10}; however, according to Theorem 3.2, we must also have $\gamma_2(u_1,v_1,1,v_2)=1$, whence the LHS is also c_0+c_{10} resulting in an equality in (3.17).

(B) The decision pair $(u_1,v_1) \notin$ C: use the results of Theorem 3.1 and an argument similar to CASE (A). □

We can also establish the following important properties for the functions Σ_2 defined in (2.4b) under Assumption 2.1 of S.ction II.

Corollary 3.5

Assuming all the strategies are PBPO, for the function Σ_2 defined in (2.4b) we have

$$\Sigma_2(u_1,1,H_0) \geq \Sigma_2(u_1,0,H_0) \qquad (3.18a)$$

$$\Sigma_2(u_1,0,H_1) \geq \Sigma_2(u_1,1,H_1) \qquad (3.18b)$$

Proof

We prove the first inequality. Using Theorem 3.3 we have

$$\Sigma_2(u_1,1,H_0) = \mathrm{E}_{v^2|H}\ J_\gamma(u_1,v_1,1,H_0) \qquad (3.19a)$$

$$\geq \mathrm{E}_{v^2|H}\ J_\gamma(u_1,v_1,0,H_0) = \Sigma_2(u_1,0,H_0) \qquad (3.19b)$$

This completes the proof. □

In other words, the assumptions in (2.6) and (2.12) of Section II are equivalent to Theorem 3.3. Due to these assumptions, the thresholds t_{u_1}, t_{v_1}, $t_{u_2}(u_1)$, and $t_{v_2}(v_1)$ are all positive quantities. It must be pointed out, however, that if we take $J_\gamma(u_1,v_1,H_j)=c_{ij}$ $(i \neq j)$ for $\forall(u_1,v_1)$:$p(u_1,v_1)=0$, i.e., for an impossible occurrence of (u_1,v_1), then we get $t_{u_1}<0$ when $p(u_1=0)=1$, and since the LR's are nonnegative quantities, this implies declaring 0 for all values of X_1. Similarly, if we take $J_\gamma(u^2,v^2,H_j)=c_0+c_{ij}$ for $\forall(u^2,v^2)$:$p(u^2,v^2)=0$, then we get $t_{u_2}(u_1)<0$ when $p(u_2=0,u_1)=1$.

(See Example 4.1 for a situation where this behavior is observed.)

We maintained earlier that the results of these theorems would save a substantial amount of computations for determining the optimal solutions. To make this more clear, note that one way of finding the optimal strategies is to first fix the supervisor's strategy γ and then solve the system of nonlinear equations in (2.15)–(2.16) or, equivalently, the system of vector equations in (2.21). Consequently, by exhausting all enumerations of the supervisor's possible strategies one can determine the optimal solution. The significance of Theorems 3.1–3.3 lies in the fact that they significantly reduce the number of possible enumerations of optimal γ.

In general, there are 2^{16} possibilities since there are four decisions, resulting in $2^4 = 16$ combinations of local decisions, and for each combination the supervisor can declare either a 0 or a 1. However, using a straightforward combinatorial calculation, it can be shown that under Theorems 3.1–3.2 the number of possibilities reduces from 2^{16} to only 1150 possibilities, i.e., a reduction of almost two orders of magnitude. As a result, the optimal solution can be obtained much faster.

IV. Computational Considerations and Numerical Results

In this section, we consider a simple example of two-step, two-detector binary hypothesis testing, and discuss two ways one can go about solving this problem. The example presented here deals with discrete and finite pdf's. Consequently, it has at least one *globally* optimal solution. All the results of the last sections are applicable here with integral signs replaced by summation signs. The purpose behind this example is to illustrate the fact that the local decision for, say, detector U at time 2 depends not only on the observations X_1 and X_2 at both times, but also on the local decision u_1 sent at time 1.

There are two facets to the decentralized sequential detection problem that we have treated: *time* and *space*. The space (hierarchy) can be broken in two ways into two interrelated minimizations: once the supervisor's decision strategy has been chosen, the problem reduces to a dynamic team decision problem. The decomposition is shown below.

$$\min_{\substack{\textit{all possible}\\ \textit{decision rules}\\ \textit{for the supervisor}}} \ \min_{\substack{\textit{all possible}\\ \textit{local strategies}}} \ E\,[J(\Gamma)]\,. \tag{4.1}$$

Conversely, once the local decision rules have been determined, so have the statistics of the "observations" available to the supervisor, and the problem from his viewpoint becomes a centralized sequential decision problem. The decomposition is shown below.

$$\min_{\substack{\textit{all possible}\\ \textit{local strategies}}} \ \min_{\substack{\textit{all possible}\\ \textit{decision strategies}\\ \textit{for the supervisor}}} \ E\,[J(\Gamma)]\,. \tag{4.2}$$

Each of these formulations adds some insight into the problem. In the former decomposition, the outer minimization is finite since there are a finite number of decisions u^2 and v^2. In general, there would be $2^{|u_1|\,|u_2|\,|v_1|\,|v_2|}$ possible decision rules for S, where the notation $|u|$ denotes the cardinality of u (which was taken to be 2 in Section II). This number grows fast as the number of detectors or the size of their decision sets increases. However, in the light of Theorems 3.1 and 3.2, these possibilities are reduced by a great deal.

Now, once the S's strategy has been selected, it is a matter of solving the system of nonlinear equations in (2.15)–(2.16) or, equivalently, the system of nonlinear vector equations in (2.21). After local strategies have been determined, the overall cost can be computed. This procedure can be repeated to exhaust all possible decision rules for S (taking into account Theorems 3.1–3.2) and choose the rule(s) which results in the smallest expected cost. The functions f_u and f_v are very complicated; fortunately (under Assumption 2.1 or, more generally, Assumption 2.2) we can perform this computation off-line.

Let us now focus our attention on discrete and finite pdf's (as in Examples 4.1 below). In this case, as was mentioned earlier, there does exist at least one *globally* optimal solution, which is given by LRT's and can be determined by finding all PBPO solutions and picking the one(s) that results in the smallest cost. For such a problem, the decomposition in (4.2) is superior. If the form of the solution were not known for the two-stage problem of Section II, all possible partitions of the X_1,X_2 and Y_1,Y_2 planes (i.e., all possible values for $p(u_1|H)$, $p(v_1|H)$, $p(u_2|u_1,H)$, and $p(v_2|v_1,H)$) would have to be considered. But we already know that the optimal solution has the form of LRT's. A straightforward calculation will show that the knowledge of this fact will result in a reduction in the number of possible decision regions from order of $2^{n_1 n_2}$ to order of $n_1^2 n_2$, where n_i is the number of points in $p(X_i|H)$. Consequently, the problem treated in Section II does not fall under any of the NP-complete categories discussed in [12] and can be solved in polynomial time.

Hence, for a problem with discrete and finite pdf's, the decomposition in (4.2) suggests the following algorithm for finding the *global* optimal solution via exhaustive enumeration:

Algorithm 4.1

A. Find LR's at time 1 and, for each detector, sort them in descending order.

B. At time 1 pick the region corresponding to the largest LR, and send 0 over that region. (This fixes the thresholds at time 1.)

  a. At time 2 compute and sort LR's for each detector.

  b. Pick the decision region corresponding to the largest LR, and send 0 over this region. (This fixes the thresholds at time 2.)

  c. Now all local strategies (i.e., all local thresholds) are specified. Find the supervisor's strategy (by, say, using Fig. 2.1) and compute the expected cost.

  d. Add the region corresponding to the next LR at time 2 to the previous decision region. Repeat part c above.

C. Add the region corresponding to the next LR at time 1 and repeat a–d above.

The regions that lead to the smallest cost compose the optimal decision rule (and there may be several of them). In the example below, we use Algorithm 4.1 to determine the global optimal strategy. (Since we deal with discrete p.d.f.'s, the possible values of the likelihood ratios will provide lower and upper limits on the thresholds.) Then, Eqs. (2.7b), (2.8b), (2.13b), and (2.14b) are used to recompute the thresholds, which must agree with the optimal decision rules found earlier. Assume samples are equally spaced starting from point 1, and define the following notation for convenience:

$$p(X|H) = \{q_1,\ldots,q_n\} \Leftrightarrow p(X=1|H) = q_1, \ldots, p(X=n|H) = q_n$$

where, obviously, $\sum_{j=1}^{n} q_j = 1$.

Example 4.1

Let U's and V's density functions be as follows:

$$p(X_1|H_0) = \{0.30, 0.40, 0.30\} \quad p(X_1|H_1) = \{0.25, 0.55, 0.20\}$$
$$p(X_2|H_0) = \{0.40, 0.10, 0.50\} \quad p(X_2|H_1) = \{0.35, 0.20, 0.45\}$$

$$p(Y_1|H_0) = \{0.10, 0.30, 0.40, 0.20\} \quad p(Y_1|H_1) = \{0.25, 0.05, 0.55, 0.15\}$$
$$p(Y_2|H_0) = \{0.30, 0.60, 0.10\} \quad p(Y_2|H_1) = \{0.20, 0.45, 0.35\}$$

Assume that X_1, X_2, Y_1, and Y_2 are mutually independent given H_0 or H_1. Take $c_{00}=c_{11}=0$, $c_{01}=c_{10}=100$, $c_0=4$, and $p_0=p_1=0.5$. Detector U's LR's at time 1 are given by

$$\lambda_{u_1}(X_1) = \{1.2000, 0.7273, 1.5000\}$$

while V's LR's are

$$\lambda_{v_1}(Y_1) = \{0.4000, 6.0000, 0.7273, 1.3333\} \cdot$$

Also, U's LR's at time 2, $\lambda_{u_2}(X_1,X_2)$ are

$$\begin{array}{c|ccc} & 1.3333 & 0.8081 & 1.6667 \\ X_2 & 0.6000 & 0.3636 & 0.7500 \\ & 1.3714 & 0.8312 & 1.7143 \\ \hline & & X_1 & \end{array}$$

while V's $\lambda_{v_2}(Y_1,Y_2)$ are

$$\begin{array}{c|cccc} & 0.1143 & 1.7143 & 0.2078 & 0.3810 \\ Y_2 & 0.5333 & 8.0000 & 0.9697 & 1.7778 \\ & 0.6000 & 9.0000 & 1.0909 & 2.0000 \\ \hline & & Y_1 & & \end{array}$$

After using the above algorithm, it turns out that there is a single optimal rule, in which U and V follow different strategies (obviously), and also each detector's strategy at time 2 depends on his decision at time 1. The following conditional probabilities of local decisions summarize the local optimal rules:

$$p(u_1{=}0|H_0) = 0.6000 \qquad p(u_1{=}0|H_1) = 0.4500$$
$$p(u_2{=}0|u_1{=}0,H_0) = 0.9000 \qquad p(u_2{=}0|u_1{=}0,H_1) = 0.8000$$
$$p(u_2{=}0|u_1{=}1,H_0) = 1.0000 \qquad p(u_2{=}0|u_1{=}1,H_1) = 1.0000$$

$$p(v_1{=}0|H_0) = 0.5000 \qquad p(v_1{=}0|H_1) = 0.2000$$
$$p(v_2{=}0|v_1{=}0,H_0) = 0.9600 \qquad p(v_2{=}0|v_1{=}0,H_1) = 0.7375$$
$$p(v_2{=}0|v_1{=}1,H_0) = 0.7200 \qquad p(v_2{=}0|v_1{=}1,H_1) = 0.4469$$

The supervisor's strategy is given by Table 4.1, where × signifies an impossible event.

For this example the expected cost is

$$E\{J_\gamma(u^2,v^2,H)\} = 32.61125$$

		u_2v_2			
		00	01	10	11
	00	0	0	0	0
u_1v_1	01	0	1	1	1
	10	0	1	×	×
	11	1	1	1	1

Table 4.1

According to this rule, U sends a 0 at time 1 if he observes a 1 or a 3, while V sends a 0 if he observes a 2 or a 4. At time 2, U sends a 0 if he gets X_2= 1 or 3 and his previous decision has been u_1=0, and he sends a 0 anyway after having sent u_1=1. V's decision at time 2 is more interesting because it depends on Y_1 as well as v_1: after having sent v_1=0, V sends v_2=0 for (Y_1,Y_2)= (2,1), (2,2), (2,3), (4,1), or (4,2); on the other hand, if he sent v_1=1 at time 1, he would send v_2=0 for (Y_1,Y_2)= (3,1) and (3,2) only.

Using the above local conditional probabilities and Table 4.1, we can compute the thresholds from Eqs. (2.7b), (2.8b), (2.13b), and (2.14b). We obtain

$$\begin{array}{rr} t_{u_1} = 1.0538 & t_{v_1} = 1.1584 \\ t_{u_2}(0) = 0.9931 & t_{v_2}(0) = 1.3750 \\ t_{u_2}(1) = -0.1094 & t_{v_2}(1) = 0.6667 \end{array}$$

As can be seen, these computed thresholds agree with the optimal strategies given above (which were computed by exhausting all possible enumerations of threshold strategies). Interestingly enough, the threshold $t_{u_2}(1)$ is negative, as it should be since at time 2, u_2=0 for all values of X_2 if u_1=1 (c.f. Table 4.1 and the discussion following Eq. (3.19b)).

V. Discussion and Concluding Remarks

We have examined in detail a two-step, two-detector decentralized hypothesis-testing problem and have discussed its straightforward extension to a decentralized multi-stage sequential detection, thereby capturing three important features of decentralized detection and decision making: (a) hierarchical structure; (b) multistage or sequential structure; (c) information rate or bandwidth reduction. We have successfully employed the Bayes criterion to find the PBPO solutions. According to each PBPO solution, under the assumption that local processors make observations independent of one another given either hypothesis, each local agent's strategy at every time instant is given by a likelihood-ratio test (LRT). It was also illustrated that each detector's threshold at each time instant depends on his past decisions (as well as his past and present observations). Thus, at the time each local agent computes a LR (Likelihood Ratio) and tests it against these thresholds, which can be computed off-line. Each detector's thresholds at a given time instant, however, are coupled with the other detector's thresholds at all times, as well as his own future thresholds.

This form of solution is not at all obvious in advance. Indeed, one might think that the local observers at, say, time 1 may well wish purposely to choose their decisions in accordance with a non-LRT policy that reveals to the supervisor more information (about, say, the local observers' conditional probability of one of the hypotheses) at time 2. As we have shown, this cannot be the case. It is therefore not unjustified to claim that the structure of the solution that has been discovered is more significant than the actual way in which one should go about determining the optimal thresholds in a specific example.

It has been pointed out recently in [12] that a large class of distributed detection problems is NP-complete, and that for the problems falling into this class computational considerations leave no alternative but to seek computable suboptimum solutions. However, not all decentralized detection problems have been shown to be NP-complete. Indeed, due to our independence assumption (Assumption 2.1 or, more generally, Assumption 2.2 of Section II) the problems treated in this paper are not NP-complete and can be computed in polynomial time. In fact, we have presented explicit computation of the decision rules and have discussed in Section III how the number of computations is reduced from an exponential order to a polynomial order.

Other issues that are worthy of consideration include examining the problem for different information patterns (e.g., allowing communication among local agents or an extension of the tandem topology discussed in [2] to multistage problems) or for the case of conditionally dependent observations.

REFERENCES

[1] Tenney, R. R. and Sandell, N. R., Jr., "Detection with Distributed Sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, July 1981, pp. 501–510.

[2] Ekchian, L. K. and Tenney, R. R., "Detection Networks," *Proc. 21st IEEE Conference on Decision and Control*, Orlando, Florida, December 8–10, 1982, pp. 686–691.

[3] Lauer, G. S. and Sandell, N. R., Jr., "Decentralized Detection given Waveform Observations," *TP-122*, ALPHATECH, Inc., Burlington, MA, February 1982.

[4] Lauer, G. S. and Sandell, N. R., Jr., "Distributed Detection with Waveform Observations: Correlated Observation Processes," *Proc. 1982 American Control Conference*, June 1982, pp. 812–819.

[5] Kushner, H. J. and Pacut, A., "A Simulation Study of a Decentralized Detection Problem," *IEEE Transactions on Automatic Control*, vol. AC-27, No. 5, October 1982, pp. 1116–1119.

[6] Teneketzis, D. and Varaiya, P., "The Decentralized Quickest Detection Problem," *IEEE Transactions on Automatic Control*, vol. AC-29, No. 7, July 1984, pp. 641–644.

[7] Teneketzis, D., "The Decentralized Wald Problem," *IEEE 1982 International Large-Scale Systems Symposium*, Virginia Beach, Oct. 1982, pp. 423–430.

[8] Teneketzis, D. and Ho, Y. C., "The Decentralized Wald Problem," preprint.

[9] Wald, A., *Sequential Analysis*, Wiley, New York, 1947.

[10] Ho, Y. C. and Chu, K. C., "Team Decision Theory and Information Structures in Optimal Control Problems–Part I," *IEEE Transactions on Automatic Control*, vol. AC-17, No. 1, February 1972, pp. 15–22.

[11] Ho, Y. C., "Team Decision Theory and Information Structures," *Proceedings of the IEEE*, vol. 68, June 1980, pp. 644–654.

[12] Tsitsiklis, J. and Athans, M., "On the Complexity of Distributed Decision Problems," *IEEE Transactions on Automatic Control*, vol. AC-30, No. 5, May 1985, pp. 440–446.

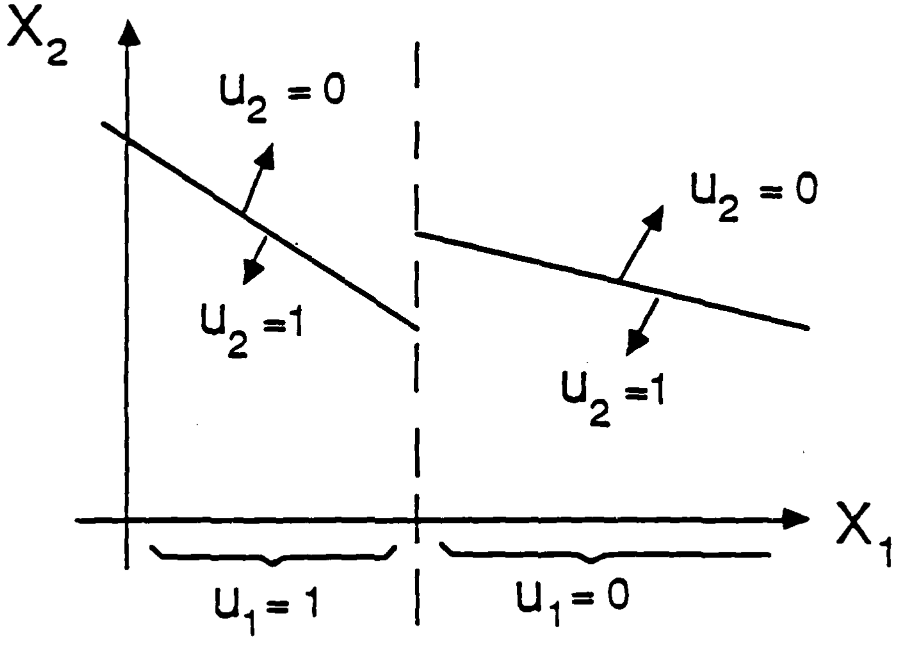


Fig. 2.2: The decision regions at time 1 and 2 for a typical example. The strategy at time 2 corresponding to each region

$$A_{u_1} \triangleq \{(X_1,X_2) : \phi_1(X_1) = u_1\}$$

is described by a likelihood-ration test. However, the strategy corresponding to the entire plane is in general not an LRT.

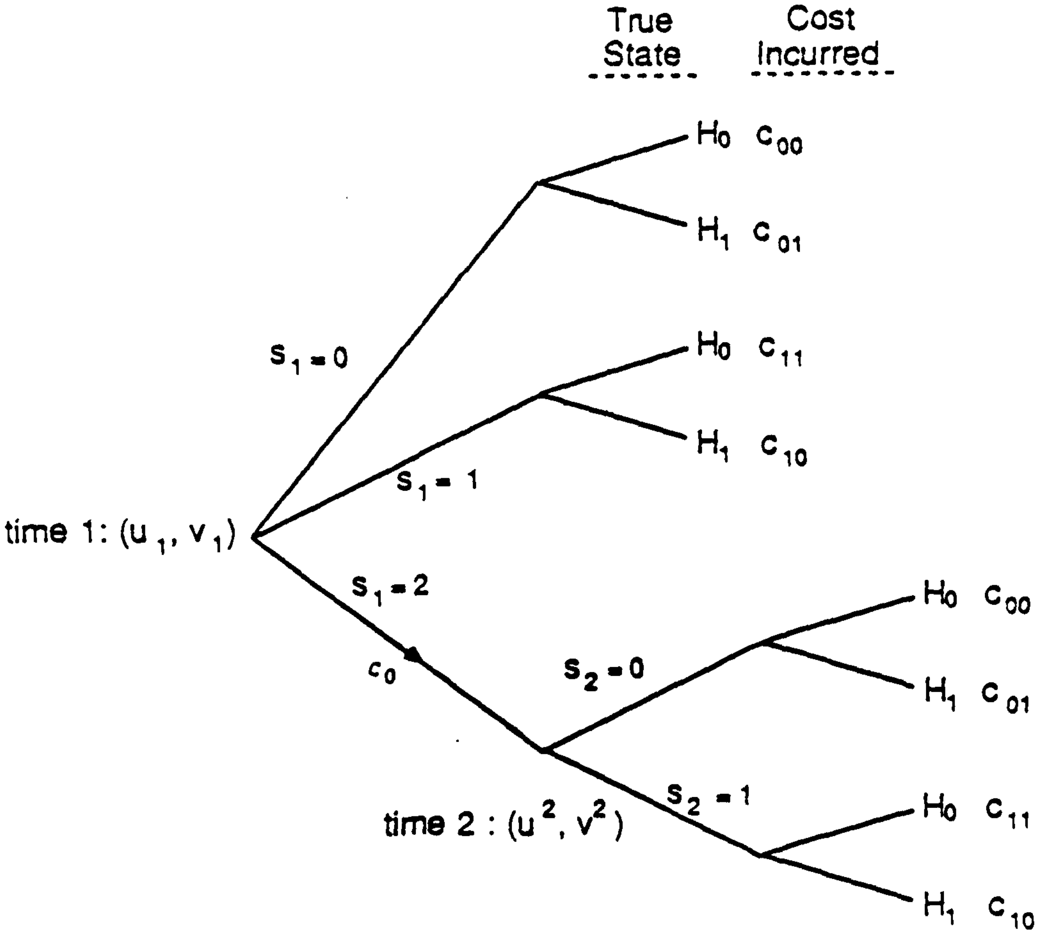


Fig 2.1: The decision tree for the two-step, two-person decentralized detection problem of Section 2.2 from the standpoint of the supervisor. At time 1 he has access to u_1 and v_1, while at time 2 he has u_1, u_2, v_1, and v_2 at his disposal.

ATE

LMED

8